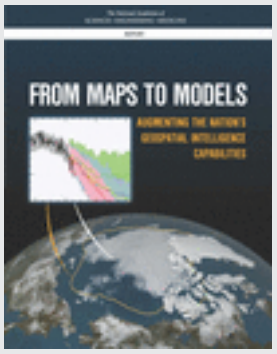


From Maps to Models: Augmenting the Nation's Geospatial Intelligence Capabilities

DETAILS

134 pages | 8.5 x 11 |
ISBN 978-0-309-44991-5 | DOI: 10.17226/23650

AUTHORS

Committee on Models of the World for the National Geospatial-Intelligence Agency; Board on Earth Sciences and Resources/Mapping Science Committee; Board on Atmospheric Sciences and Climate; Division on Earth and Life Studies; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

FROM MAPS TO MODELS

AUGMENTING THE NATION'S GEOSPATIAL INTELLIGENCE CAPABILITIES

Committee on Models of the World for the National Geospatial-Intelligence Agency

Board on Earth Sciences and Resources/Mapping Science Committee
Board on Atmospheric Sciences and Climate
Division on Earth and Life Studies

Board on Mathematical Sciences and Their Applications
Division on Engineering and Physical Sciences

A Report of

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by the National Geospatial-Intelligence Agency under Contract No. HM017713C0002. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-44991-5
International Standard Book Number-10: 0-309-44991-X
Digital Object Identifier: 10.17226/23650

Additional copies of this publication are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Cover: (Upper) Past and projected Arctic sea ice decline, 1890–2090. Observed values (black line) are for 1953–2012. Projected values (colored lines, 2010–2090) represent different future policies for regulating CO_2, from very high emissions (red line) to substantially declining emissions after 2020 (green line). SOURCE: Modified from Stroeve et al. (2012). (Lower) Arctic sea ice summertime minimum on September 10, 2016 (white) compared to 1981-2010 average minimum (gold line). SOURCE: NASA Goddard's Scientific Visualization Studio/C. Starr.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2016. *From Maps to Models: Augmenting the Nation's Geospatial Intelligence Capabilities*. Washington, DC: The National Academies Press. doi: 10.17226/23650.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Reports document the evidence-based consensus of an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and committee deliberations. Reports are peer reviewed and are approved by the National Academies of Sciences, Engineering, and Medicine.

Proceedings chronicle the presentations and discussions at a workshop, symposium, or other convening event. The statements and opinions contained in proceedings are those of the participants and have not been endorsed by other participants, the planning committee, or the National Academies of Sciences, Engineering, and Medicine.

For information about other products and activities of the National Academies, please visit nationalacademies.org/whatwedo.

COMMITTEE ON MODELS OF THE WORLD FOR THE NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY

DAVID M. HIGDON, Chair, Virginia Tech, Arlington, Virginia
ROBERT L. AXTELL, George Mason University, Fairfax, Virginia
VENKATRAMANI BALAJI, Princeton University/Geophysical Fluid Dynamics Laboratory, New Jersey
LAWRENCE E. BUJA, National Center for Atmospheric Research, Boulder, Colorado
KATHERINE V. CALVIN, Pacific Northwest National Laboratory/Joint Global Change Research Institute, College Park, Maryland
KATHLEEN M. CARLEY, Carnegie Mellon University, Pittsburgh, Pennsylvania
REBECCA CASTAÑO, Jet Propulsion Laboratory, Pasadena, California
RONALD R. COIFMAN, Yale University, New Haven, Connecticut
OMAR GHATTAS, The University of Texas at Austin
JAMES A. HANSEN, Naval Research Laboratory, Monterey, California
ANNA M. MICHALAK, Carnegie Institution for Science, Stanford, California
SHASHI SHEKHAR, University of Minnesota, Minneapolis
SHAOWEN WANG, University of Illinois at Urbana-Champaign

THE NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE STAFF

ANNE M. LINN, Study Director, Board on Earth Sciences and Resources
EDWARD J. DUNLEA, Senior Program Manager, Board on Atmospheric Sciences and Climate
SCOTT T. WEIDMAN, Director, Board on Mathematical Sciences and Their Applications
ERIC J. EDKIN, Senior Program Assistant, Board on Earth Sciences and Resources

Acknowledgments

The report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their participation in the review of this report:

Erica Briscoe, Georgia Tech Research Institute, Smyrna
Paul K. Davis, RAND Corporation, Santa Monica, California
Auroop R. Ganguly, Northeastern University, Boston, Massachusetts
Michael F. Goodchild, University of California, Santa Barbara
Alexander H. Levis, George Mason University, Fairfax, Virginia
Richard M. Medina, University of Utah, Salt Lake City
Guillermo Sapiro, Duke University, Durham, North Carolina
Cyrus Shahabi, University of Southern California, Los Angeles
Jery R. Stedinger, Cornell University, Ithaca, New York

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions nor did they see the final draft of the report before its release. The review of this report was overseen by George M. Hornberger, Vanderbilt University, and Keith C. Clarke, University of California, Santa Barbara, who were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

The committee would like to thank the following individuals who shared their expertise with the committee through presentations and discussions: Richard Berg, MITRE; Nadya Bliss, Arizona State University; Michael Chatman, Pacific Disaster Center; Edward Cope, NGA; Dolores Derrington, MITRE; Peter Douchette, Integrity Applications, Inc.; Kerri Dugan, NGA; David Gauthier, NGA; Matt Hancher, Google; Dirk Helbing, ETH Zurich;

Kevin Krystopolski, NGA; Eric Lance, Arizona State University; David Lawrence, NCAR; Michael Lenihan, NGA; Jon Miller, Arizona State University; Sherry Olsen, MITRE; Peter Overton, MITRE; H. Greg Smith, NGA; Monica Smith, NGA; Tim Stearns, Stanford University; and Dave White, Arizona State University.

Contents

Summary

The United States faces numerous, varied, and evolving threats to national security, including terrorism, scarcity and disruption of food and water supplies, extreme weather events, and regional conflicts around the world. Effectively managing these threats requires intelligence that not only assesses what is happening now, but that also anticipates potential future threats. The National Geospatial-Intelligence Agency (NGA) is responsible for providing geospatial intelligence on other countries—assessing where exactly something is, what it is, and why it is important—in support of national security, disaster response, and humanitarian assistance. NGA's approach today relies heavily on imagery analysis and mapping, which provide an assessment of current and past conditions. However, augmenting that approach with a strong modeling capability would enable NGA to also anticipate and explore future outcomes.

A model is a simplified representation of a real-world system that is used to extract explainable insights about the system, predict future outcomes, or explore what might happen under plausible what-if scenarios. In this report, a model means a mathematical or numerical model that can be run on a computer. Such models use data and/or theory to specify inputs (e.g., initial conditions, boundary conditions, and model parameters) to produce an output.

At the request of NGA, the National Academies of Sciences, Engineering, and Medicine established a committee to describe types of models and analytical methods used to understand real-world systems; to determine what would be required to make these models and methods useful for geospatial intelligence; and to identify supporting research and development for NGA (see Box S.1). The report provides examples of models that have been used to help answer the sorts of questions NGA might ask; describes how to go about a model-based investigation, using example questions to illustrate how NGA might think about choices and tradeoffs; and discusses models and methods that are relevant to NGA's mission.

MODEL-BASED INVESTIGATIONS

Models do not stand alone, but rather exist in an environment that includes the available data, methods for analysis and model assessment, computational and data infrastructure, and people skilled in developing, tailoring, and running models and interpreting their output. A model-based investigation begins by formulating the key questions to be answered. The questions drive the choice of models, analytical methods, data, and computational resources as well as how these pieces will be combined to generate results with the necessary speed and accuracy. Using existing models or model output speeds the investigation and reduces its cost. If appropriate models or model

BOX S.1 Committee Tasks

1. Identify types of mathematical, numerical, and statistical models and spatiotemporal analytical methods (e.g., coupled models, inverse models, agent-based models, machine learning, and statistical inference) used to understand complex adaptive systems, such as those found in the natural or built environment, and in health, political, social, or economic systems.
2. Describe the potential relevance of these models and methods to geospatial intelligence.
3. Describe the current state of the art in the models and methods relevant to geospatial intelligence, including factors such as the features and scales captured by the model, accuracy, reliability, predictability, uncertainty characterization, and computational requirements.
4. Determine what would be required to make these models and methods useful for geospatial intelligence, considering issues such as adaptability of the model for other purposes, availability of data, interoperability, and computational issues.
5. Identify NGA research and development necessary to adapt, populate, link, analyze, and maintain the models and methods for geospatial intelligence purposes.

output do not exist, the investigators have two choices: develop a new model from scratch or combine existing subsystem models into a new model. A combined model has the potential to capture the behavior of the larger system, as long as both the processes in the subsystem models and their connections are appropriately represented.

Once chosen, the models or model products are incorporated into an analysis, typically involving data and computation, to connect the model to the real-world system. A variety of analytical methods may be used in the investigation, such as methods to preprocess data before ingesting them into the model, to update the model state as new observations accrue, to calibrate model parameters, to combine subsystem models, to determine what new data to collect, or, crucially, to assess the credibility and uncertainty of the results. A model differs from the real-world system it seeks to represent for several reasons, including omitted or inadequate representation of system processes, errors and uncertainties in the data used as input to the models, and coding errors in the models. Any model-based investigation must assess the impact of these uncertainties on our inferences about the real-world system, and communicate this uncertainty to decision makers.

The demands of the analysis dictate the computer infrastructure needed. Some models and methods can be run on a laptop. High-performance computing is generally needed for computationally intensive models that require large numbers of processors (e.g., thousands), dedicated communication between processors, and large volumes of data (e.g., gigabytes to terabytes) in memory and storage (e.g., climate models). Data-intensive computing is used for the analysis of massive amounts of data (e.g., terabytes to petabytes), which is dominated by data processing tasks. In these cases, computation and data manipulation must be divided into parallel tasks that can operate on separate pieces of the data, with minimal communication between these separate tasks (e.g., text mining).

MODELS AND METHODS FOR NGA

Given the breadth of national security and humanitarian challenges under NGA's purview, it could be argued that dozens of models and analysis methods, each with important variants, are potentially relevant to NGA. The study was unclassified in its entirety, and so the committee used NGA's mission, the special characteristics of geospatial data, and two example intelligence scenarios provided by NGA to guide its selection of relevant models

and methods. Below is a discussion of models and methods that seem particularly relevant to NGA, illustrated by the following NGA intelligence scenario:

> China needs to find more water to meet agriculture and energy demands, but major dam projects (e.g., Three Gorges Dam) have displaced large populations against their will. Broad intelligence questions include *How do agriculture and energy production and consumption change over time? How and where will populations, including rural communities, shift?*

Types of Models and Methods Relevant to NGA

The first two tasks of the committee were to identify types of models and methods used to understand complex systems and describe their relevance to NGA (see Box S.1). NGA's mission is to produce geospatial intelligence, which is defined as the exploitation and analysis of imagery and geospatial information to describe, assess, and visually depict physical features and geographically referenced activities on Earth. To extend this mission to the modeling realm, NGA will need models of human behavior (activities), set within an environmental context (physical features), and techniques for integrating and analyzing geospatial (geographically referenced) data. This analysis will be influenced by the spatial, space-time, or network structure that is present in geospatial data. Key characteristics of such data include autocorrelation (e.g., the properties of nearby locations tend to be similar) and spatial heterogeneity (i.e., the phenomenon being modeled varies with location).

These needs place a premium on the following types of models and methods:

- ***Models of physical processes that affect human activities.*** Such models use theory-driven equations to describe environmental (atmospheric, oceanic, hydrologic, and geologic) systems and to predict their future behavior. For the NGA intelligence scenario, for example, a large-scale physical process model of the hydrologic system in China could be used to predict surface flow, subsurface flow, and abundance of water under different water diversion scenarios.
- ***Social system models of human behavior in a geospatial context.*** Particularly relevant models include system-level models of influences and flows between stocks, and agent-based models, which use defined rules of behavior among individuals to study emergent population behavior. These models support what-if reasoning, such as how different scenarios are likely to unfold. For example, social system model scenarios could be constructed to assess of how affected populations are likely to respond to dam construction and involuntary migration in China.
- ***Models of combined physical and social systems.*** Physical and social subsystem processes that depend on one another may be coupled to understand their interactions at different locations. For example, weather, wave, and pirate behavior models have been combined with shipping patterns to help predict where and when pirates might attack. More complicated economic, resource, and energy interactions and feedbacks may be captured in integrated assessment models. For example, integrated assessment models could be used to examine the complex interactions among water, agriculture, and energy production and consumption in China.
- ***Inverse methods to infer uncertain model parameters from measurements of the real-world system.*** These methods combine computational models with real-world observations to constrain model parameters (e.g., physical constants, initial conditions, boundary conditions, and system states) so the model better represents the real-world system and therefore produces more reliable results. For example, inverse methods could be used to estimate or constrain key model parameters of a large-scale hydrologic model (e.g., spatially varying permeability, flow rates, and evaporation) to produce plausible predictions of water availability as a function of location throughout China. Inverse methods must be supplied with problem-specific information and tailored to the model being used.

• ***Spatial statistics, data mining, and machine learning to discover trends, patterns, and associations.*** These methods use data-driven empirical models and methods to understand what diverse observations reveal about the system. For example, empirical models could be used to detect changes in water availability arising from policy decisions on dam building, agriculture, and coal production in China. Empirical approaches are often customized to the type of data being analyzed (e.g., spatiotemporal data, images, and text) and the available computational environment.

• ***Spatial network analysis to examine how patterns of relations affect behavior at the individual to state level.*** Network analysis models are graphical or statistical models that explain behavior among people related by friendship, financial, cultural, or other ties. Spatial network models, which combine spatial and network reasoning, explain how patterns of relations affect behavior. Network models have been used to characterize the actors in covert groups and their relationships, to assess the potential for various kinds of attacks, and to evaluate changes and trends in dynamic geopolitical environments. They are often used in scenario descriptions and what-if analyses, and temporal trends in network data are also used for prediction. For example, social media analytics could be used to determine which neighborhoods and which groups in China are likely to strongly resist migration required by dam building.

State of the Art

The third task of the committee was to describe the current state of the art in relevant models and methods, including features and scales, accuracy, reliability, predictability, uncertainty characterization, and computational requirements (see Box S.1). These factors are applicable to a particular model or method in a particular setting, but not to the broad classes described above. Consequently, the report provides ranges or examples for the various factors called out in Task 3.

The state of the art influences which models and methods NGA can adapt in the near term and which require additional research and development. In general, physical process models and data-driven methods (inverse, empirical, or network analysis) are relatively mature, but will have to be adapted for geospatial intelligence purposes. Examples of useful developments include downscaling techniques, which would facilitate the application of NGA's remotely sensed data in problems on smaller, more policy-relevant scales, and empirical methods that can handle specialized model structure (e.g., space-time dependence) and features of geospatial data. Social system and coupled physical–social system models can support what-if reasoning and scenario development, which are useful for many NGA applications. However, fundamental research is needed to improve understanding of human behavior and to make the models easier to develop.

Increasing the Usefulness of Models and Methods to NGA

The fourth task of the committee was to determine what would be required to make relevant models and methods useful for geospatial intelligence (see Box S.1). What particular models and methods to develop or adapt for geospatial intelligence purposes depends primarily on their utility for the investigations at hand, but also on the difficulty of developing or adapting the models and methods, the availability of software and code, the level of training support, and the knowledge and experience of NGA analysts. For example, NGA must often respond quickly to an emerging national security threat or a natural disaster, and so a simple model that is relatively quick to develop, adapt, or run may be more useful than a more comprehensive model that takes months or longer to develop. When a more sophisticated model is needed, NGA could leverage the considerable expertise of modeling groups at universities, national laboratories, federal agencies, and private companies. Actions NGA can take to develop or adapt the models and analysis methods described above are summarized below.

In-house development or adaptation. Data-driven models and analysis methods are amenable to near-term development, because NGA analysts already have some relevant knowledge and experience, the methodology is established, and software, tools, and training support are available. In particular, NGA's experience with spatial and temporal analysis provides a foundation for developing or adapting spatial statistics, data mining, and machine learning methods. Methods that are especially promising for NGA include (a) Bayesian hierarchical models that link spatially connected data, data at different levels of resolution and aggregation, or disparate data sources; (b) clustering and other unsupervised and deep learning methods for finding structure in large volumes of data that are too much to analyze with a human in the loop; and (c) methods for detecting change footprints and spatial hotspots and anomalies. In addition, NGA's growing emphasis in human geography provides a foundation for developing network analysis models to examine how patterns of relations affect behavior.

For both types of analyses, the basic methods are well established, and software and user support (e.g., textbooks, conferences, and special short courses) are readily available. However, some additional development and training are required to adapt these methods for geospatial data and NGA use cases. In addition, software and algorithms for data-intensive computing will likely have to be developed for spatial statistics and spatial data mining methods. Training in network or spatial network analysis could be offered at the NGA College or obtained from university-based programs.

Collaborations. NGA will need partners to help develop, adapt, and use more sophisticated models and methods (e.g., process models, coupled models, agent-based models, inverse methods, and spatial network models) as well as geospatial models that are not well supported by cutting-edge computational infrastructure. A substantial part of any group's capability in sophisticated modeling is learned through partnerships, apprenticeships, and collaborations. Such collaboration could take many forms, including being a partner in the team developing a model or extending its use to other applications, a user of a team's model or method, or a user of the resulting data products. Regardless, NGA will need to identify domain experts who can design models or scenarios relevant to NGA, run the model, interpret the results, or help NGA find useful existing model output. To use these models or model results effectively, NGA will need to understand their strengths and limitations for the geospatial intelligence task at hand.

Finding partners for NGA modeling efforts will not be trivial because of the classified nature of the work, the wide and changing variety of experts needed, and the need to nurture long-term relationships. Models of complex systems are typically developed by multidisciplinary teams with in-depth knowledge and experience in the scientific disciplines and computational capabilities relevant for the task at hand. However, bringing together diverse experts, who would learn from each other in the context of NGA's priorities, could contribute to major breakthroughs in NGA-relevant problems. Major research universities, as well as organizations for which NGA has established relationships (e.g., defense and intelligence agencies, national laboratories, private-sector contractors, and NGA centers of academic excellence in geospatial science) may be a starting point for finding experts and modeling teams for NGA modeling efforts.

NGA-Funded Research and Development

The fifth task of the committee was to identify areas that could benefit from NGA-funded research and development (see Box S.1). A host of investments in research and development could strengthen NGA's modeling capabilities in the years ahead. Where to focus these investments depends on what models and analysis methods are proving most useful for geospatial intelligence. Potential research areas concern extending the use of existing models to NGA-relevant situations, improving understanding of human behavior, reducing the time required to develop, test, and run models, and developing methodologies tailored to NGA-relevant models and data, as described below.

Extending the use of models to NGA-relevant situations. NGA investigations will likely make new demands on models, using them in settings for which they have not been originally designed, or at least not thoroughly tested. Examples include precise, near-real-time wind, wave, or weather predictions to support troop deployment, disaster relief, or dispersion and damage estimates from the release of hazardous materials in urban environments. In addition, such models may need to be combined with social system models to help decision makers prepare for social unrest, disruption, or migration. Substantial research is required to adapt physical process models, social system models, and combined physical–social system models to deal more reliably with these less common settings.

Improving understanding of human behavior. Social system models are only beginning to surpass expert judgment. Advancing their development, and the development of combined physical–social system models, requires fundamental research to improve understanding of human behavior. Promising areas of research include studies aimed at understanding how human behavior is constrained or enabled by the geography of the natural and built environment, including how geographic factors influence the development of social networks and communications among actors, and how cognitive biases influence the perception of space.

Speeding model development, testing, and run time. Intelligence questions are often time sensitive, and so research advances that speed up model development, testing, or run time could prove beneficial to NGA. Model development could be sped up through research aimed at facilitating the combination of existing subsystem models for NGA investigations. Model development and run time could be decreased through research and development of accurate reduced models that use coarser, simpler, or fewer representations of processes than computationally intensive models. Developing simulation testbeds could aid all of these efforts and also facilitate assessments of model accuracy and speed.

Methodological research and development tailored to NGA-relevant models. The models developed or adapted for NGA purposes will have to be accompanied by customized methods that combine these models with data. Methodology for inversion, exploration of plausible outcomes or scenarios, quantification of prediction uncertainties, and model assessment will be required to bring model-based results more in line with available measurements. Such methodology is particularly needed for social system and physical–social system models. Possible directions in this area include development of inverse methods for constraining the plausible states of social system models to be consistent with data, and development of methods for formal verification and validation of model results against NGA-relevant benchmarks and test cases. In addition, research on how to adapt existing inverse methods to integrate the diverse forms of data that NGA collects and uses (e.g., satellite, sensor, geospatial, and open source) would be beneficial for all types of models.

Methodological research and development tailored to NGA data sources and needs. Research could advance the development of empirical methodology by tailoring it to data and use cases found in NGA applications. Examples include developing methods to combine disparate data or results from different approaches and to more accurately represent their uncertainty in support of inference and decision making, and methods to cope with data that have spatial, temporal, and network structure (e.g., to assess the activities of a terrorist cell over time). A related need is for sentiment-mining techniques that characterize the sentiment in a document based on the geographic and network features of the community, such as location, local events, language, structure of local groups, and tendency to self-identify in networks. Research is also needed to develop algorithms for promising spatial and spatiotemporal methods to efficiently leverage the processing and data storage capabilities present in advanced data-intensive computational architectures.

OVERARCHING CONCLUSIONS

Overall, the committee concludes that model-based investigations would provide NGA with a powerful means to search for spatial and temporal patterns; make inferences from those patterns; predict future political, economic, and military threats to the United States; and evaluate options and consequences around the world. However, models are not the optimal tool for every geospatial intelligence problem, such as when time is too short to carry out a model-based investigation or when a lack of data or uncertainties about key processes make it difficult to obtain useful insights on the real-world system. Consequently, mapping and imagery analysis will continue to be important for geospatial intelligence.

NGA can begin to develop or adapt some data-driven models now, given analysts' long experience with spatial and temporal analysis and the availability of supporting software and tools. Developing a more sophisticated modeling and analysis capability will likely take many years, because NGA will need new knowledge, skills, techniques, and workflows; interactions with external modeling groups; additional sources of data; increased computational capabilities; and a change to a modeling mindset. NGA need not build this capability for cutting-edge or complex models developed, maintained, or supported by a large research community (e.g., climate models). Indeed, working with external modeling groups to use their models, model output, and analysis methods for geospatial intelligence purposes would stretch resources and help analysts gain knowledge and expertise in modeling and analysis methods. As their experience grows, NGA analysts will be able to take on progressively more complex model-based investigations in the classified world.

1

Why Models?

We live in a world of numerous, varied, and evolving threats to national security, including terrorism, proliferation of weapons of mass destruction, scarcity and disruption of food and water supplies, extreme weather events, natural disasters, and regional conflicts and disputes (Clapper, 2015; NIC, 2012). Many of these threats have an important spatial component, and understanding where exactly something is, what it is, and why it is important is the job of the National Geospatial-Intelligence Agency (NGA). NGA analysts evaluate imagery and remote sensing and other data on the land, ocean, and atmosphere to create geospatial intelligence in support of national security, disaster response, and humanitarian assistance (see Figure 1.1).

FIGURE 1.1 An NGA analyst at work. Analysts acquire, process, and analyze imagery and other geospatial information from a variety of classified and open sources and deliver information products and services to policy and decision makers. SOURCE: http://gcn.com/articles/2014/11/20/nga-map-of-the-world.aspx.

NGA grew out of a merger of several defense and intelligence organizations that provided mapping, charting, imagery analysis, and geospatial information services.[1] Initially called the National Imagery and Mapping Agency, the name was changed to NGA in 2004 to mark the emergence of the discipline of geospatial intelligence. The term geospatial intelligence is defined as "the exploitation and analysis of imagery and geospatial information to describe, assess, and visually depict physical features and geographically referenced activities on the Earth."[2] Geospatial information includes collections of data items, each referenced to a location on the surface of Earth. It may describe places at multiple scales (e.g., countries, provinces, counties, and neighborhoods); physical objects (e.g., buildings and vehicles); and spatiotemporal activities (e.g., movement of people), relationships (e.g., networks among people, institutions, and places), and patterns (e.g., hotspots and co-occurrences).[3] These attributes are relevant to both mapping and modeling.

NGA has developed some models that characterize geophysical features and phenomena (e.g., land and seafloor terrain, and magnetic and gravity fields [see Figure 1.2]) and, increasingly, the human geography (e.g., distribution, alliances, and hostilities among ethnic groups across a country). However, NGA envisions developing a broader modeling and analysis capability that will enable analysts to search for spatial and temporal patterns; make inferences from those patterns; predict future political, economic, and military threats to the United States; and evaluate options and consequences around the world. This added capability would be used to better inform policy choices, action plans, and contingency strategies that depend on geospatial intelligence.

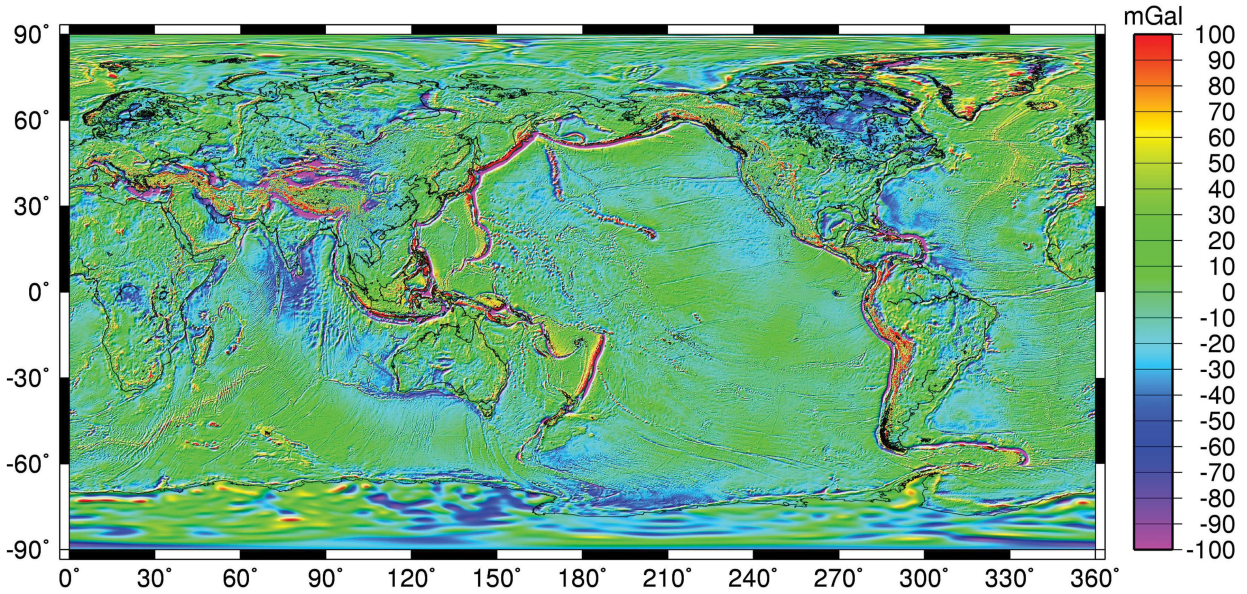


FIGURE 1.2 Global map of free air gravity anomalies (difference between measured and reference gravity values, corrected for elevation, in milligals [mGal]), computed from NGA's Earth Gravitational Model 2008. Features with relatively high mass (e.g., mountains and subduction zones) have high measured gravity values relative to the reference value, and a positive gravity anomaly (warm colors). NGA uses these data to support navigation systems, mapping, and surveys. SOURCE: NGA, http://earth-info.nga.mil/GandG/wgs84/gravitymod/egm2008/anomalies_dov.html.

[1] NGA Reference Chronology, available at https://www.nga.mil/About/History/Pages/Educational-Resources.aspx.

[2] 10 USC § 467.

[3] Richard M. Medina and George F. Hepner, A note on the state of geography and geospatial intelligence, https://www.nga.mil/MediaRoom/News/Pages/StateofGeographyandGEOINT.aspx.

At the request of NGA, the National Academies of Sciences, Engineering, and Medicine established a committee to describe mathematical, numerical, and statistical models and spatiotemporal analytical methods used to understand real-world systems; to determine what would be required to make these models and methods useful for geospatial intelligence; and to identify supporting research and development for NGA (see Box 1.1). NGA offered two pieces of guidance for addressing these tasks. First, a wide variety of public-domain, proprietary, and classified models and methods are likely relevant, and NGA asked the committee to focus on those in the public domain, which have largely been developed by university researchers. Second, the study was unclassified in its entirety, and little information was available on NGA's current capabilities and future needs. NGA provided an overview of the agency and geospatial intelligence, as well as two example intelligence scenarios (see Box 1.2). The committee used this information, national security threats with a geospatial component (e.g., Clapper, 2015; NIC, 2012), and members' experience with NGA and other defense agencies to guide its selection of potentially relevant models and methods, and its suggestions for NGA research and development needs.

BOX 1.1 Committee Tasks

1. Identify types of mathematical, numerical, and statistical models and spatiotemporal analytical methods (e.g., coupled models, inverse models, agent-based models, machine learning, and statistical inference) used to understand complex adaptive systems, such as those found in the natural or built environment, and in health, political, social, or economic systems.
2. Describe the potential relevance of these models and methods to geospatial intelligence.
3. Describe the current state of the art in the models and methods relevant to geospatial intelligence, including factors such as the features and scales captured by the model, accuracy, reliability, predictability, uncertainty characterization, and computational requirements.
4. Determine what would be required to make these models and methods useful for geospatial intelligence, considering issues such as adaptability of the model for other purposes, availability of data, interoperability, and computational issues.
5. Identify NGA research and development necessary to adapt, populate, link, analyze, and maintain the models and methods for geospatial intelligence purposes.

DEFINING MODELS

Generally, a model is a simplified representation of a real-world system that enables investigation of system properties and behaviors. In this report, a "model" means a mathematical or numerical model that can be run on a computer. It typically requires inputs that specify initial conditions, boundary conditions, and/or model parameters to produce an output. A model could be as simple as a linear mapping from input to output ($x \rightarrow ax + b$), or as complex as a climate model that includes multiple processes operating at multiple and temporal and spatial scales, evolves the state of Earth's ocean and atmosphere over centuries, and runs on a supercomputer. Such models are used to extract explainable insights about the system, to enable prediction of future outcomes, or to aid in decision making by simulating multiple "what-if" scenarios for consideration.

Models tend to fall somewhere on the spectrum ranging from data driven to theory driven. Data-driven empirical models (e.g., regression and classification models from the fields of statistics and machine learning) are designed to ingest data and efficiently estimate model parameters, capturing dependencies and correlations between system features. Theory-based process models encode causal connections between states, entities, or subsystems, describing system evolution and/or response to changing conditions. The appropriate level of empiricism depends

BOX 1.2 Example Intelligence Scenarios Provided by NGA

Megacities

Megacities (>10 million people) are key instruments of social and economic development, and they affect the future prosperity and stability of the world. A broad intelligence question is *How will worldwide urbanization trends affect regional political, economic, and security environments?* Intelligence tasks include the following:

- What change indicators from geospatial intelligence and other sources can be best measured to assess these conditions or events?
- What kinds of models are most suitable for predicting or forecasting the changes in urbanization that could trigger political, economic, or security problems?
- Can models of today's megacities help us understand national security issues in current and future megacities?

Chinese Water Transfer Project

Seventy percent of water in China is used for agriculture and 20 percent is used for energy production. China must find enough water to develop new coal reserves. However, the Three Gorges Dam project (see Figure 1.2.1) displaced 330,000 people against their will. Broad intelligence questions include *How do agriculture and energy production and consumption change over time? How and where will populations, including rural communities, shift?* Intelligence tasks include the following:

- What change indicators from geospatial intelligence and other sources can be best measured to assess these questions?
- What kinds of models are most suitable for predicting or forecasting how a project like this could affect political, economic, environmental, or national security problems?
- Can models help us understand the impacts of current and future massive government-funded public works projects?

FIGURE 1.2.1 Three Gorges Dam. SOURCE: Courtesy of Keith Clarke, University of California, Santa Barbara.

SOURCE: NGA.

on a number of factors, such as the availability of system observations, the computational demands of the model, the maturity of the theory, and the needs of the model-based investigation.

MODELS FOR UNDERSTANDING REAL-WORLD SYSTEMS

The notion of building models of real-world systems has a long history. Among the first predictable, recurring phenomena humans tried to model were astronomical in nature, including the diurnal cycle, the passing of the seasons, and the movement of planets in the night sky. Early astronomers, including Aryabhata (born in India in 476) and Tycho Brahe (born in Denmark in 1546), carried out the immensely complex computations by hand. The advent of digital computing in the mid-20th century offered a way to perform massive computations based on mathematical formulas, and revolutionized modeling capabilities. Weather and climate modeling is a classic example of how digital computing transformed an entire field (see Box 1.3).

The current situation in social system modeling is not unlike weather forecasting 35 years ago, when computational methods first began surpassing expert judgement (Edwards, 2010). For example, the Federal Reserve Board now relies on both experts (board members) and models (e.g., FRB/US) to estimate the near-term prospects for the U.S. economy, which are then used to decide on policy. The performance of both the experts and the models has been checkered. In the 2007–2009 financial crisis, the FRB/US model correctly predicted the steep drop in housing prices, but substantially underestimated the rise in unemployment. Expert predictions of unemployment were also too optimistic, even while the crisis was unfolding. A variety of computationally sophisticated social system modeling efforts are now under way in economics (e.g., Delli Gatti et al., 2011), finance (e.g., Geanakoplos et al., 2012), and other policy realms (e.g., state stability and epidemics). Consequently, prospects for these domains to follow the trajectory of weather model development are strong.

THE MODEL-BASED INVESTIGATION PROCESS

An investigation of a real-world system is typically an iterative process, framed by the central tasks of (1) identifying key questions to be addressed, (2) scoping the investigation, (3) exploiting models to make inferences about the key questions, (4) assessing the model-based analyses, and (5) revising any of these steps as necessary.

Identifying Key Questions or Features to Explore

A model investigation is focused on particular questions or features of a system. These questions drive how models will be exploited, ensuring that the investigation yields information relevant to the users. Different user requirements may lead to fundamentally different models of the same phenomenon. Consequently, it is important to choose the questions carefully. For example, the broad intelligence question in NGA's megacities scenario (*How will worldwide urbanization trends affect regional political, economic, and security environments?* [see Box 1.2]) likely encompasses several model-based investigations on different aspects of the issue (e.g., the probability and impact of heat waves and the impact of changing demographics on crime).

A related task is to ask what value modeling will add in addressing the key questions. In some cases, the complexity of the system relative to the sophistication and fidelity of available models, or the paucity of relevant observations of the system, makes the effort nearly impossible. Choosing investigations where quantitative modeling is likely to shed light on the key questions is crucial.

BOX 1.3 A Brief History of Weather and Climate Modeling

Early weather forecasts were made by meteorologists, who laboriously constructed contour maps of temperature and other variables by hand, using experience and tradecraft to fill in between sparse observations (Edwards, 2010). Predigital attempts at numerical forecasting were computationally laborious for humans, and the noteworthy attempt by Richardson (1922) failed because the limitations of numerical methods were not recognized. Digital computing enabled forecasts to be made using equations of fluid flow. The first attempts to build forward weather models were made in 1950 (Platzman, 1979). By 1956, it was clear that model-based forecasting was going to be a revolutionary advance over the prior heuristic methods (Phillips, 1956), and a number of centers around the world began to make both heuristic and numerical weather predictions. For some years, weather forecasts by human analysts remained superior to those generated by a computer. However, the skill of computer forecasts has improved steadily over time, and modeling is now the primary approach for making operational weather forecasts.

The same models, run for long periods, were also used to test Arrhenius' 19th-century conjecture that adding CO_2 to the atmosphere might cause the planet to warm. However, it soon became apparent that the model had to include both the atmosphere and the oceans, which take up heat. The first coupled atmosphere–ocean model (Manabe and Bryan, 1969) has been recognized as a milestone in scientific computing (Ruttimann, 2006). In 1975, Manabe and Wetherald ran a coupled climate model to equilibrium with two different levels of CO_2 (one at present-day levels, the other at doubled CO_2) and found distinct warming at the Earth's surface (see Figure 1.3.1). Based in part on results from two models, the "Charney Report" confirmed that planetary-scale warming was likely on the timescale of decades, and suggested that this warming could have an influence on society (NRC, 1979).

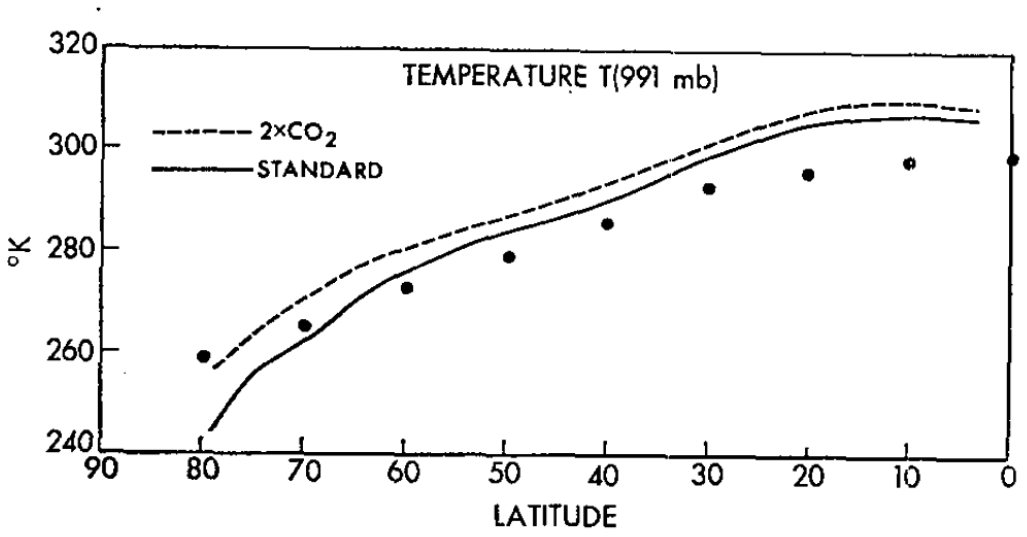


FIGURE 1.3.1 Equilibrium climate response to doubled CO_2 (dashed line) shows warming at the planetary surface. The dots in this figure represent globally averaged surface air temperature observations as a function of latitude, from Oort and Rasmusson (1970). SOURCE: Manabe and Wetherald (1975). © American Meteorological Society. Used with permission.

These results led to the establishment of international programs, including the Intergovernmental Panel on Climate Change (IPCC), which was formed in 1988. Today, IPCC assessment reports provide the scientific basis for policy decisions on climate change at global to regional scales. The enterprise involves some of the largest supercomputers in the world, globally coordinated modeling experiments (known as Coupled Model Intercomparison Projects), and contributions from thousands of scientists and engineers. In fact, the collection of activities—from data collection, to scheduled runs of large models to generate future projections of surface fields such as temperature and precipitation (e.g., Figure 1.3.2), to the dissemination of information products to the public—has been described as a vast machine (Edwards, 2010).

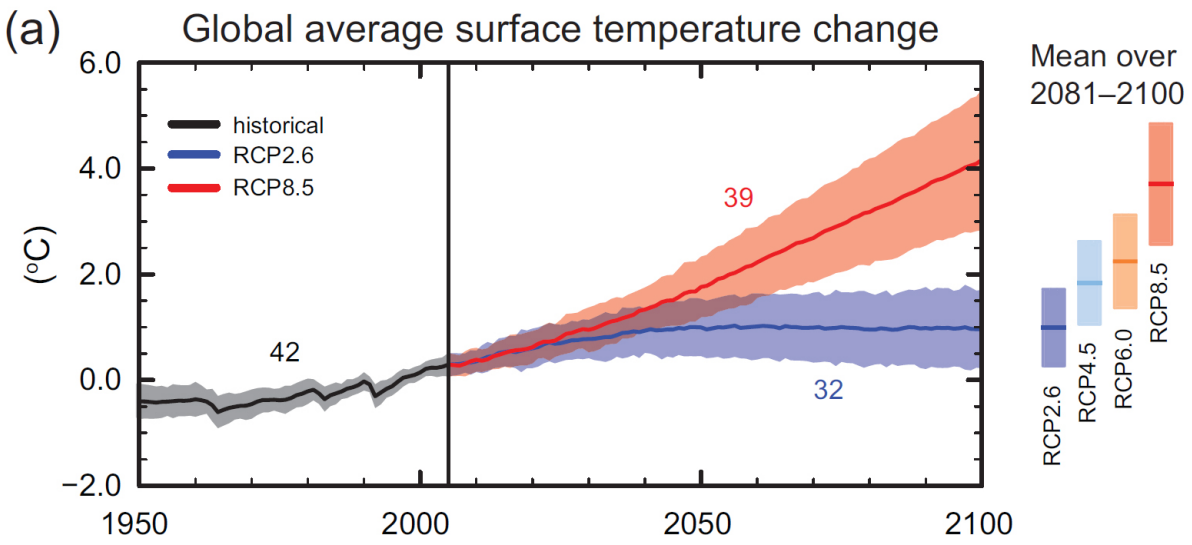


FIGURE 1.3.2 Historical (black line) and simulated surface temperature change (red and blue lines, relative to 1986–2005) under different future policies for regulating CO_2, ranging from very low climate forcing (RCP2.6) to climate stabilization (RCP4.5 and PCP6) to very high greenhouse gas emissions (RCP8.5). SOURCE: Figure SPM.7(a) from IPCC (2013).

Scoping and Planning the Investigation

The real-world systems associated with NGA's example intelligence scenarios contain a large number of heterogeneous subsystems (e.g., geophysical, environmental, cultural, and economic), and the models used to investigate them are computationally challenging, are data rich, and require an arsenal of analytic approaches. Scoping includes determining what models, data, analytical methods, subject-matter expertise, computational resources, and other resources are required for the investigation. For example, different models have different capabilities, such as their ability to capture particular aspects of reality, and these capabilities must be compared with the characteristics needed from the analysis to decide on a modeling approach. Scoping is intimately connected to plans for combining these ingredients to provide useful information about the key questions. The speed and accuracy of required results, the availability and maturity of existing models, and the availability and completeness of data also affect scoping.

Exploiting Models

A model-based investigation may involve developing a new model, setting up and running an existing model, combining or coupling existing models, and/or analyzing results from previous model runs or analyses. The models or model products are then incorporated into an analysis, typically involving data and computation, to connect the model to the real-world system under investigation and to produce insights, predictions (and uncertainties), or plausible outcomes for some features of the system. For NGA investigations, the analysis will be influenced by the spatial, space-time, or network structure that is present in geospatial data. Key characteristics of such data include autocorrelation (e.g., the properties of nearby locations tend to be similar) and spatial heterogeneity (i.e., the phenomenon being modeled varies with location).

Model exploitation is at the heart of any investigation, and it is explored more broadly in Chapter 3. It is worth pointing out here that the strengths and limitations of the models used are inherited by the larger investigation.

Assessing the Credibility of Model-Based Insights

A model differs from the real-world system it seeks to represent for several reasons, including omitted or inadequate representation of system processes, errors and uncertainties in the data used as input to the models, and coding errors in the models. Any model-based investigation must assess the impact of these uncertainties on our inferences about the real-world system and communicate this uncertainty to decision makers. Ideally, this assessment takes place throughout the development of the model, addressing the question of whether the behavior of a model sufficiently matches the behavior of the real-world system to address the key question(s) of the investigation. With mature models and sufficient data, the assessment process often covers the adequacy of the model's representation to the real-world system (validation) and the quantification of model uncertainty. For models with exploratory and extrapolative elements or insufficient data for comparison, assessment processes may be more qualitative, particularly in situations where uncertainty is large and difficult to quantify. Evaluation processes might also involve comparing predictions or outcomes from different models, especially if they use different conceptual modeling approaches.

Making Revisions

Information and insight gained throughout the investigation is used to revise any of the elements mentioned above—including the formulation of key questions, the application of resources, the computing infrastructure used, and the data collected—to improve the model and the quality of analysis results. For example, the empirical elements of a model may be upgraded as new data become available. Thus, the model representation itself may evolve in response to new data.

STRENGTHS AND LIMITATIONS OF MODELS

Models simplify relationships and omit or aggregate some features and processes of the real-world systems they are representing, depending on the key questions of the investigation. This abstraction is precisely what makes models immensely useful, because it reduces the number of processes that may be acting or changing and thus provides clearer insight on the most important aspects of the system. In addition, a model can be explored in ways the real-world system cannot. For example, sensitivity studies can be carried out to determine how varying key model inputs causes model outputs to change. Simulation-based experiments can be carried out to assess the model's response to particular input forcings or management strategies.

Models have led to increased understanding and produced accurate predictions in a wide variety of physical

systems, including weather (Simmons and Hollingsworth, 2002; see Box 1.3), nuclear physics (Hendricks et al., 2008), and astrophysics (Fryxell et al., 2000). Even a simple, empirical model can lend insight on a complicated system, as illustrated by Francis Galton's use of correlation and regression to understand features of heritability in the late 1800s (Stigler, 1986). On the other hand, models did not anticipate the accumulation of local influences and external forcings that caused the collapse of the Grand Banks fishery in 1992 (McGuire, 1997) or the financial crisis in 2008 (Taylor, 2009).

Models also have limitations in capturing important behaviors of complex systems. Many of the systems of interest to NGA have a large number of interrelated and autonomous subsystems, and both these subsystems and the larger system can adapt. In such complex adaptive systems, complexity can be thought of as the potential for emergent behavior to appear and for small changes to have unforeseen and potentially enormous consequences (Holland, 1992; Lansing, 2003). Such systems often exhibit nonlinear behavior, and the same action at different points in the system's history may have different results.

The gap between model and reality can be substantial when the real-world system itself is evolving toward a new regime that is unlike previous experience. For models that have been tuned to perform well in comparison to past data, there is a danger that the model and reality will diverge in a new regime, even if the model has adequately tracked the system in the past. Model-based predictions of climate change embody this concept. Many climate models can accurately simulate the global temperature record, but their predictions of warming for the 21st century can differ by a factor of 2 or 3, depending on how the models handle aerosols and greenhouse gases, which contribute to warming in different ways (Kiehl, 2007). Hence, accuracy in reproducing past conditions is no guarantee of accuracy in projecting future conditions, since the combined (and potentially nonlinear) effects of two different forcings may not be fully realized in the historical data alone. A change in regime also affects data-driven models, which are trained to learn correlations and patterns from existing data.

In addition, prediction accuracy declines over longer timescales because of complex or chaotic system behavior. The limits of predictability in fully specified, completely deterministic systems have been demonstrated by Lorenz (1963). Chaotic systems might also contain multiple regimes and exhibit transitions (abrupt changes from one equilibrium to another), tipping points (changes from one stable regime to another), hysteresis, and path-dependent behaviors. Figure 1.3 shows a trajectory in phase space which appears to orbit around one center of attraction, but a very small displacement can shift the trajectory to an orbit around an entirely different center of attraction. A small error in the initial condition can thus shift the system into a different equilibrium. Although such systems pose challenges for prediction and understanding, they also exhibit tendencies and behaviors that can be explored with models.

Model initial conditions, parameters, or forcings can be perturbed to develop a range of potential trajectories, which can be helpful for exploring possible future outcomes, testing analysis approaches, decision making, or training. For example, potential trajectories are often combined with the perceived cost or benefit of each trajectory and the user's attitude about decision making under uncertainty (e.g., Howard, 1968; Keeney, 1982; Raiffa, 1968). Model-based trajectories have been used to guide decision making in a variety of application areas, including climate mitigation (e.g., Drouet et al., 2015), smallpox interventions (e.g., Ferguson et al., 2003), and terrorism response (e.g., Rosoff and von Winterfeldt, 2007).

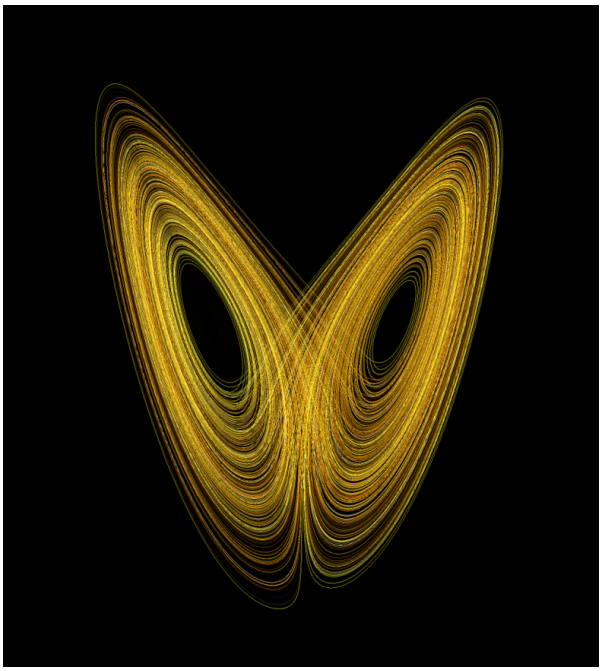


FIGURE 1.3 The Lorenz butterfly, emblematic image of chaos. The figure depicts the strange attractor that can unpredictably and abruptly switch between multiple regimes. SOURCE: Wikimedia.

Because most decision makers or other users have not been involved in the modeling, it is particularly important to communicate the strengths, weaknesses, and overall level of uncertainty of the model-based investigation. As cautioned by Chesshire and Surrey (1975):

> Because of the mathematical power of the computer, the predictions of computer models tend to become imbued with a spurious accuracy transcending the assumptions on which they are based. Even if the modeler himself is aware of the limitations of the model and does not have a messianic faith in its predictions, the layman and the policymaker are usually incapable of challenging the computer predictions. Depending upon the modeler's credibility, a dangerous situation can arise in which computation becomes a substitute for understanding about problems and relationships, and a substitute rather than an aid for policy choices.

Even the results from simple models are subject to misinterpretation if their many assumptions and uncertainties are not conveyed clearly. Communicating uncertainty of scientific results remains an active area of research (e.g., Fischoff and Davis, 2014; Morgan and Henrion, 1990; NRC, 2006, 2007a).

ORGANIZATION OF THE REPORT

This report describes models and analytical methods that are potentially relevant to NGA and discusses how they could be developed, used, or adapted for geospatial intelligence purposes. Chapter 2 illustrates the power of models and methods to tackle issues of potential interest to NGA. Chapter 3 describes the components of a model-based investigation, including the models themselves, the data linking the model to the real-world system,

methods for analysis and model assessment, and the computational infrastructure. Example questions are posed to illustrate how NGA could think about these components. Chapter 4 covers models and methods likely to be of particular interest to NGA, the state of the art in these models and methods, ways to make them useful to NGA, and research and development that could help NGA adapt, use, and maintain the models for geospatial intelligence purposes. Additional detail on methods for combining models and on computation appears in Appendixes A and B, respectively. Biographical sketches of committee members are given in Appendix C, and acronyms and abbreviations appear in Appendix D.

2

Illustrative Models

A wide variety of model-based investigations have been carried out successfully in areas of potential interest to the National Geospatial-Intelligence Agency (NGA). This chapter describes five examples relevant to NGA's mission:

1. Pirate attacks. NGA provides support (human geography information, imagery analysis, cartography) to the Navy and other defense agencies fighting piracy at sea, as well as warnings to commercial ship captains.[1]
2. Housing bubbles. Financial stresses are a key component of NGA's megacities scenario (see Box 1.2).
3. Disease outbreaks. NGA provides support for humanitarian assistance activities, such as base maps showing cultural places and communication, power, and transportation infrastructure in affected African countries during the outbreak of the Ebola virus disease.[2]
4. Tensions in the Middle East. Human geography is a core geospatial intelligence capability, and it features in a variety of analyses of regional conflicts.[3]
5. Food and water scarcity. Competing food, energy, and water demands are a driver for NGA's Chinese water transfer scenario (see Box 1.2).

These examples are intended to illustrate how models have been used to help answer the sorts of questions NGA might ask, not to detail all of the elements and complexities of the particular model-based investigations. Each example includes representative forensics, tactical, or strategic questions that might guide a model-based investigation. Forensics questions guide investigations aimed at understanding why a particular event occurred. Tactical questions are aimed at supporting short-term decisions, such as the deployment of resources. Strategic questions are aimed at longer-term decisions and planning that could benefit from projections or scenarios of possible futures. These types of investigations are often related. For example, the results of forensics and tactical models may inform the development of strategic models, and strategic models may supply scenarios or initial conditions for tactical models.

[1] See https://www.nga.mil/About/History/Pages/A-Generation-of-Geospatial-Intelligence.aspx.
[2] See NGA Pathfinder Magazine 12(4):20-24, 2014 [online]. Available at https://www.nga.mil/MediaRoom/Pathfinder/Pages/Archive.aspx.
[3] See NGA Pathfinder Magazine 14(1):10-11, 2016 [online]. Available at https://www.nga.mil/MediaRoom/Pathfinder/Pages/Archive.aspx.

EXAMPLE 1: PIRATE ATTACKS

Piracy around the Horn of Africa (Indian Ocean) costs the United States up to $16 billion per year (Hansen et al., 2011) and also hampers humanitarian efforts and endangers lives. Counterpiracy forces must allocate their limited assets (about 30 vessels) to protect several million square miles of ocean from thousands of pirates. For example, more than 2,000 pirates operated in Somalia alone in 2011 (Kirk, 2011). These pirates generally attack large commercial ships using small vessels that are highly vulnerable to rough ocean conditions.

Tactical Question: How should ships and aircraft be deployed to best protect commercial shipping from piracy?

A key element of mitigation is to determine where the pirates are and when they will attack so that counterpiracy forces can be in the right place at the right time. The three conditions that must be met for a pirate attack to occur are (1) the presence of a vulnerable ship, (2) the presence of pirates, and (3) mild weather conditions (generally waves must be less than 1 m for the pirates to be able to board the ship under attack). Combining models of weather, commercial shipping, and recent pirate attacks, Naval Research Laboratory researchers developed a model for predicting where the pirates will be and when they are most likely to strike (Hansen et al., 2011). Later, the model was enhanced with historical weather data and an agent-based model of pirate behavior to create a Pirate Attack Risk Surface product (Slootmaker et al., 2013; see Figure 2.1). The product is validated using measures of reliability (e.g., events forecast with a 30 percent chance of occurring should occur about 30 percent of the time), and measures of sharpness (e.g., the region with a predicted high probability for pirate attack is smaller than the region covered in a climatological prediction). The results indicate that the model produces sharper forecasts than climatology, and that pirate attacks are more likely to occur in areas where the predicted probability of attack is high.

These models and products are continually evolving as shipping patterns and pirate behavior change. For example, the behavior model can be updated as pirates adapt to antipiracy efforts or change their capabilities. This product is briefed daily to the senior leadership of U.S. Naval Forces Central Command to support decision making on deploying antipiracy assets.

Strategic Question: What conditions lead to a greater risk of piracy?

This question calls for an examination of conditions that lead to increased piracy, and many different types of subsystem models could be used to address different aspects of the problem. For example, climate models could provide information on changing environmental conditions, which could then be input to models of water and food availability. Infrastructure models could examine the logistics associated with distribution of food, water, and aid. Societal models could reveal the dynamics of tribes, religion, and centralized government and the ability of these groups to enhance or reduce the conditions that make piracy attractive. Similar models could also be used to understand the creation, growth, and sustainment of organizations necessary to recruit, train, and supply pirates. The policies of insurance companies and countries that influence the availability of funds for ransom could also be factored into the analyses.

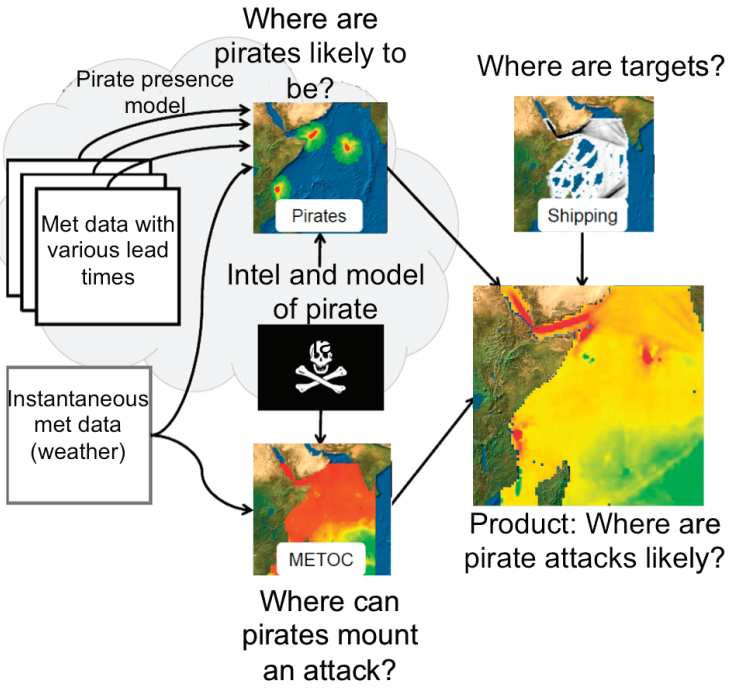


FIGURE 2.1 Elements used in the implementation of the pirate attack risk surface product. NOTE: METOC = meteorological and oceanographic. SOURCE: Slootmaker et al. (2013).

EXAMPLE 2: HOUSING BUBBLES

Economic bubbles are characterized by a boom and then a collapse of asset prices, accompanied by uncertainty about the fundamental value of the assets. The run-up is often fueled initially by demand and then largely by speculation. Because it is difficult to assess fundamental asset value, it is difficult to diagnose a bubble until it bursts.

Speculative asset price bubbles can be traced back as far as ancient Rome (e.g., Chancellor, 1999). A more recent example is the U.S. housing bubble of 2005–2008. Housing prices peaked in early 2006, started to decline in 2006 and 2007, and reached new lows in 2012. The credit crisis resulting from the burst of the housing bubble is considered one of the primary causes of the 2007–2009 economic recession in the United States (Holt, 2009), and it had ripple effects across the globe.

Strategic Question: How do various financial policies affect housing bubbles, such as the U.S. housing bubble of 2005–2008?

Recent advances in computing, data availability, and modeling have made it possible to combine microdata with agent-based modeling to study how the behavior of individual households and banks can produce bubbles at the aggregate level. In agent-based models, the agents (in this case, households and banks) interact directly with one another and with their economic environment, following autonomous decision rules. By studying the behavior of house prices in the periods leading up to and after the crisis, and investigating the importance of various factors that are amenable to policy interventions (e.g., central bank interest rate policies, leverage limits, creditworthiness limits, removal of refinancing restrictions, write-downs of mortgage principal, and subsidies to first-time homeowners), it may be possible to identify policies that could help prevent a bubble and promote financial stability.

Geanakoplos et al. (2012) combined several independent data sets on household behavior in the Washington,

DC, metropolitan area and created an agent-based model of the 2005–2008 housing bubble and its aftermath. As shown in Figure 2.2, the model recreated the bubble given the actual interest rates and loan down payment (leverage) requirements (upper left). It also suggested that fixed interest rates would reduce but not eliminate the bubble (upper right), and that freezing the down payment required would significantly reduce the boom and eliminate the bust (lower panels).

An interesting feature of the U.S. housing bubble is its spatial variability—a handful of cities (e.g., Las Vegas, Miami, and Phoenix) experienced a relatively severe bubble, while other cities (e.g., Dallas) experienced very little rise in house prices. In Canada, Toronto experienced a large increase in house prices but no decline, and thus no bubble. More recently, large price increases in housing in London caused the Bank of England to express concern that that city could be in the middle of a bubble (*The Guardian*, July 15, 2014) and led to tightened lending standards for mortgages. Models of the type described above can be useful in determining whether price increases in specific cities present significant bubble risk.

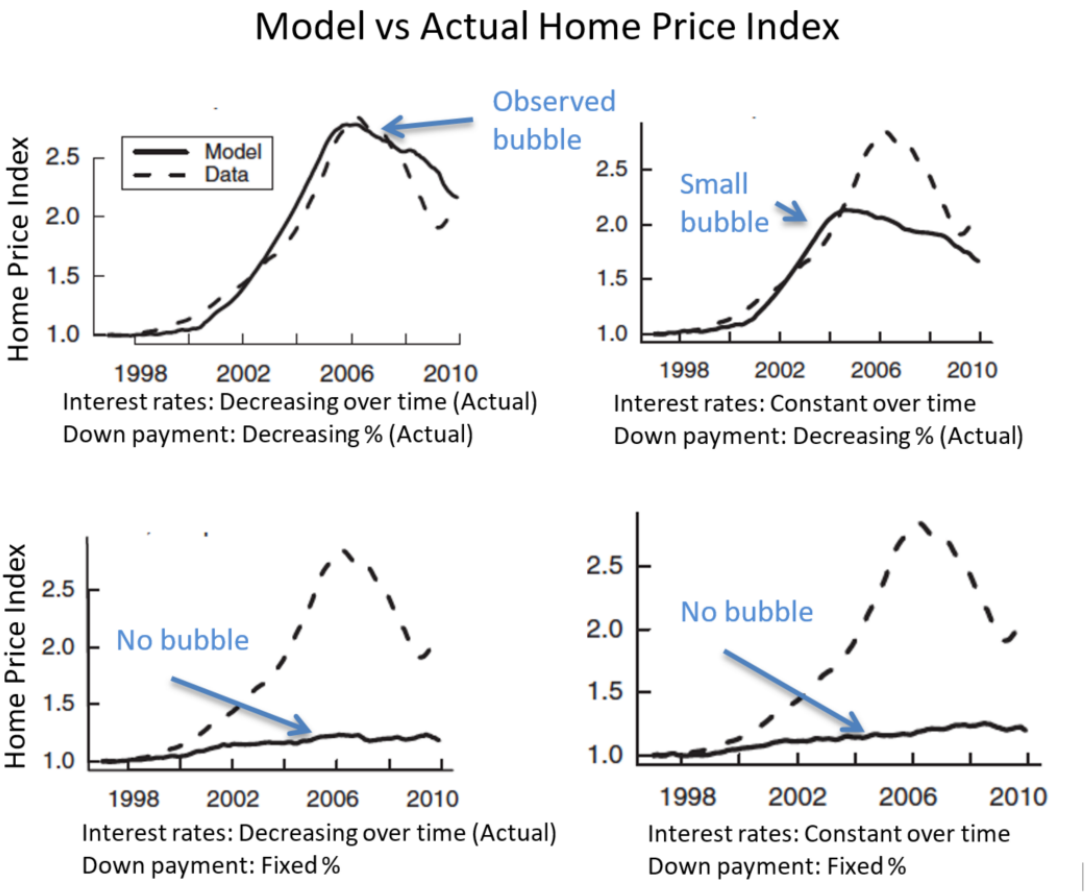


FIGURE 2.2 Housing bubble model for Washington, DC. The plots show the Case-Shiller housing index normalized to 1 at the start of the time period. (Upper left) The model correctly recreated the bubble using the actual historical interest rates and loan down payment (leverage) requirements. (Upper right) Freezing interest rates, but allowing the loan down payment requirements to vary historically, led to a smaller bubble. (Lower left) Freezing the loan down payment requirements while allowing the interest rates to vary historically yielded no bubble. (Lower right) If both the leverage and interest rates were held constant over time, no bubble appeared. SOURCE: Geanakoplos et al. (2012).

EXAMPLE 3: DISEASE OUTBREAKS

In the 1800s, cholera epidemics were common in cities and were thought to be caused by noxious air emanating from rotting organic matter. John Snow mapped cholera death locations in the SoHo neighborhood of London during the 1854 outbreak and noticed a geographic concentration (now called a hotspot) of deaths with a water pump at the center (Snow, 1855). Snow persuaded city leaders to shut down the water pump, and the disease subsided. Snow's map is an early example of how simple spatial context (distance to water pump) can be used to refine a disease model. Today, modelers consider spatial, spatiotemporal, temporal, and social contexts to study patterns of disease, as exemplified in the 2014 outbreak of Legionnaires' disease in Portugal. Models are also used to address strategic and tactical questions related to the spread of disease, such as malaria. These examples are discussed below.

Legionnaires' Disease

People contract Legionnaires' disease by breathing in aerosols (small droplets of water in the air) that have been contaminated with Legionella bacteria. The bacteria grow in warm freshwater environments, such as hot tubs, showers, or air-conditioning units. Legionnaires' disease usually strikes individuals, not groups. However, in November 2014, 18 patients with Legionnaires' disease were admitted to two hospitals in the municipality of Vila Franca de Xira, Lisbon, within 24 hours. Within hours, the Portuguese Ministries of Health and Environment convened a multidisciplinary task force to control the outbreak (Shivaji et al., 2014).

Tactical Question: Once an outbreak of Legionnaires' disease occurs, how many cases are expected?

An early estimate of the expected number of cases can help public health officials deploy resources. Egan et al. (2011) used data from a number of historical outbreaks of Legionnaires' disease to develop an empirical model that convolved an infection time distribution (representing the aerosolized release) and an incubation period distribution. In Portugal, the date of onset of symptoms reported by patients was used to estimate the period of exposure to the contaminated aerosols, and thus the date of the contaminant release (Shivaji et al., 2014). Using these results, health officials were able to estimate expected cases and revise their anticipated demand for ventilator systems, one element of the contingency plan for hospitals in the region.

Tactical and Forensic Question: What is the source of the bacteria causing the outbreak?

Legionnaires' bacteria do not spread from person to person. The disease outbreak is usually halted following closure or thorough cleaning of the source, and so rapid identification of the source is critical for containing the outbreak. In Portugal, analysis of the spatial distribution of the residences of affected individuals indicated that 90 percent lived within 3 km of a wet cooling system (Shivaji et al., 2014; see Figure 2.3). The task force hypothesized that the most likely source of infection was aerosolized release from one or more of these systems, and it used Egan et al.'s model to identify cases whose pattern of illness did not fit with this hypothesis. The patient who developed the earliest symptoms was found to have been involved in maintenance in the wet cooling system towers in the two weeks before onset of symptoms.

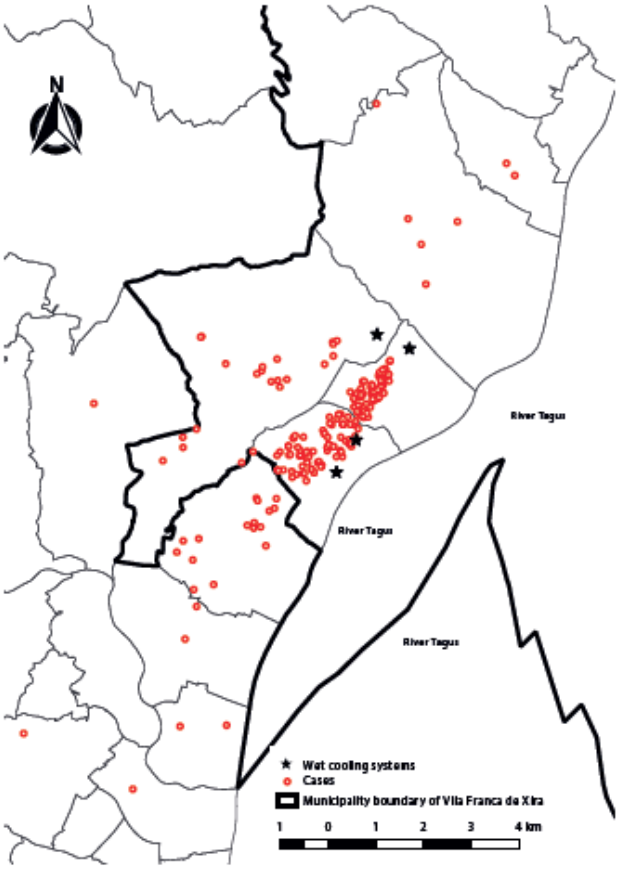


FIGURE 2.3 Spatial distribution of cases of Legionnaires' disease in Villa Franca de Xira, Portugal, in November 2014. Red dots indicate residences of patients and stars indicate the locations of wet cooling systems. SOURCE: Shivaji et al. (2014).

Malaria

Strategic Question: What is the expected impact of global warming on malaria?

Malaria is an infectious disease with one of the largest health burdens worldwide. In 2015, there were more than 200 million malaria cases worldwide, and malaria killed more than 400,000 people, most of them children under 5 years old in Africa.[4] The impact is so great in African countries that a high incidence of malaria has been associated with reduced economic growth (Gallup and Sachs, 2001). Malaria is transmitted by mosquitos, which require specific environmental conditions to live. Thus, changes in temperature and precipitation affect the distribution and transmission of the disease.

The potential impact of a changing climate on malaria incidence has been studied through process-based and empirical models. Process-based models use equations that express the relationship between climatic variables and biological parameters (e.g., mosquito breeding, survival, and biting rates; Plasmodium parasite incubation rates). They are useful for analyzing the climate sensitivity of malaria risk, although at scales much larger than mosquito habitats. Process model results suggest that a small increase in temperature can greatly increase malaria

[4]See http://www.who.int/features/factfiles/malaria/en.

transmission (Caminade et al., 2014; Pascual and Bouma, 2009), potentially allowing expansion of malaria to higher-altitude areas and increasing the seasonal duration of malaria in some endemic areas. Empirical models are useful for defining the distribution limits of the disease, although they do not reveal the mechanisms driving climate sensitivity of malaria risk. Empirical models using recorded cases and observed temperature have confirmed the occurrence of malaria in higher altitudes with increased temperatures (Siraj et al., 2014).

Current models generally overestimate the impact of climate change on global malaria distribution. In fact, despite significant increases in temperature, the range of malaria has decreased considerably over the past century, largely due to economic development and disease control efforts (Gething et al., 2010; see Figure 2.4). Climate-induced effects are more consistent with the observed changes over a few regions of Africa and South America (Caminade et al., 2014), where socioeconomic conditions have resulted in limited interventions. Some areas where mitigation efforts have either been less intensive or been reduced have seen a resurgence of malaria (WHO, 2012). Thus, inclusion of socioeconomic factors with climate impacts is critical when using models to support policy decisions.

Tactical Question: Can likely outbreaks of malaria be forecast with enough lead time to position resources to mitigate the effects?

Early warning systems for outbreaks of malaria currently rely on monitoring the caseloads of malaria infections. Once a threshold is crossed in number of cases reported, alerts of an epidemic are issued, and resources are repositioned accordingly. As noted above, transmission and infection rates of malaria have been linked to rainfall and temperature variation. For example, epidemics usually emerge in Botswana in the 2 months following the November–February rainy season. This means that longer-term forecasts of temperature and rainfall could potentially lead to better warning systems for likely malaria outbreaks.

Early research (Thomson et al., 2006) has linked the emergence of epidemics to December–February seasonally averaged rainfall and sea-surface temperatures, which enables forecasts of epidemic risk with a month lead time. A growing body of research suggests that integrating monthly to seasonal forecasts of sea-surface temperatures, cumulative rainfall, and temperature variability could extend malaria early warnings 4 to 6 months (Jones and Morse, 2012; Lauderdale et al., 2014; MacLeod et al., 2015; Thomson et al., 2006; Tompkins and Di Giuseppe, 2015; see also NASEM, 2016a). Forecasting malaria outbreaks 2 to 6 months in advance would allow tactical positioning of malaria prevention programs (e.g., spraying pesticides) and resources (e.g., distributing prophylactic drug therapies for target groups during the malaria season, and shifting medical resources to ensure timely and effective care for infected individuals; Myers et al., 2000). Improved modeling capability could help to inform these types of resource allocation choices.

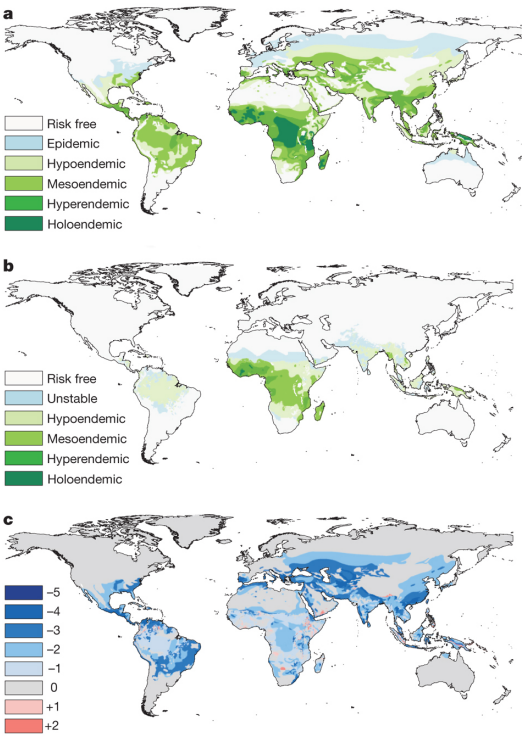


FIGURE 2.4 Global malaria endemicity (a) in 1900, before mitigation, and (b) in 2007, with mitigation measures in place. The 2007 map was developed using surveys of Plasmodium parasite rates and model-based geostatistics for the spatial prediction of malaria endemicity (Hay et al., 2009). (c) Comparison of malaria endemicity between 1900 and 2007. Negative values indicate a decrease in endemicity. SOURCE: Gething et al. (2010). Reprinted by permission from Macmillan Publishers Ltd.: Nature, © 2010.

EXAMPLE 4: TENSIONS IN THE MIDDLE EAST

Tensions in the Middle East present a wide range of intelligence community challenges, such as characterizing the actors (e.g., ethnic groups, stakeholders, and emergent leaders) in covert groups and their relationships; assessing the potential for various types of attacks; and evaluating changes and trends in dynamic geopolitical environments. The data needed to tackle these challenges are so distributed and the situation is changing so rapidly that open-source data and rapid assessments are often key. With recent advances in language technology, social network analysis, dynamic network analysis, and agent-based models, it is now possible to develop models from open-source and other data in a semiautomated human-in-the-loop fashion in a matter of days (Carley et al., 2012a; see also Chapter 4).

Strategic Question: Is it possible to predict a violent revolutionary uprising, such as the Arab Spring?

Predictive analysis is needed to position military forces, first responders, or humanitarian aid workers to mitigate or minimize the negative impacts of an uprising. Although such analysis is challenging, a piece of the puzzle can be addressed: Is there any signal in the open-source data that such an uprising is likely? Joseph et al. (2014) explored this question using newspaper articles and agent-based dynamic network simulation for 16 countries associated with the Arab Spring. Their analysis produced forecasts for each country, expressed as monthly trends in the change in two beliefs: the need for revolution (protests, riots, and government overthrows) and the need for

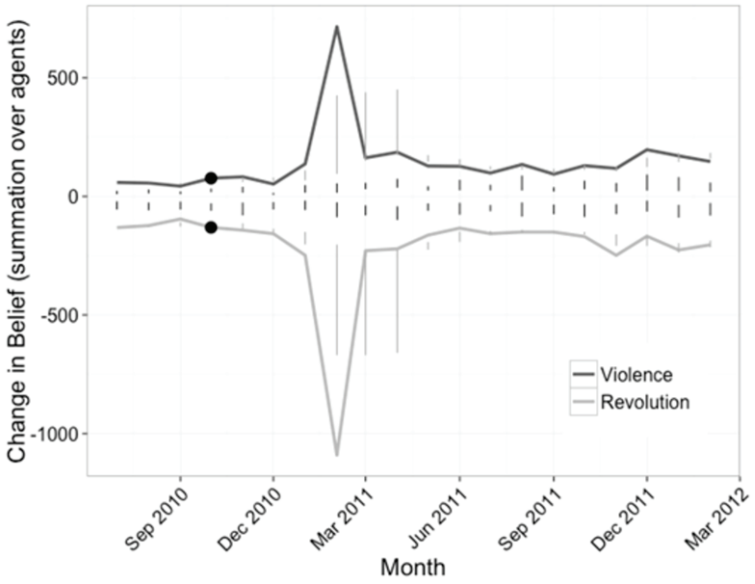
violence. The results were validated against historical data to determine whether the predicted trend first peaked in the months precipitating the violent revolutionary event, or whether there were no peaks or other eventualities. Figure 2.5 shows how the model fared against the key initiating events in these countries. The model correctly predicted revolution would not occur in the six countries only tangentially involved in the Arab Spring (low and sometimes decreasing revolutionary activity). In addition, the model correctly predicted revolution in three out of four countries whose governments were overthrown (high levels of revolutionary activity and high violence) and correctly predicted lower levels of revolutionary activity in six countries whose governments are still in place. An increase in relevant topics appearing in newspapers, a shift to more violent topics, and an increasingly dense network of actors were key to predictive skill.

It is important to recognize that these are very high-level predictions. For example, the model predicted that the peak in violence and revolutionary activity would occur in Tunisia prior to the other countries. It did not, however, predict the trigger event—the self-immolation of Mohamed Bouazizi, a street vendor, in response to his treatment by authorities.

Forensics Question: Was the 2012 attack on the Benghazi consulate planned or the result of a spontaneous anti-U.S. protest?

The same historical data and network models used in the Joseph et al. (2014) study predicted the September 2012 attack on the Cairo embassy, but not that on the Benghazi consulate. Network analytics can help show why. News and Twitter posts as the events were unfolding showed that the two events had very different footprints in the open-source data (Carley et al., 2014). A set of models for each attack was developed in less than 4 hours and then updated every 4 hours. Figure 2.6 shows a Twitter retweet social network (top) and a topic network based on hashtags (bottom) for Benghazi. Retweet networks are a common way of tracking tweeters of disproportional influence and the flow of a "story." Hashtag networks are a common way of tracking the connection among ideas, and examining how the "story" changes across groups, locations, and time. These models showed that in Benghazi, and less so in Cairo, the key actors in the retweet network were news agencies. When the Benghazi consulate was attacked, news articles quickly appeared questioning the role of the movie The Innocence of Muslims on the attack. However, there was no mention of the movie prior to the event, and all references to the movie after the attack began with news agencies. Twitter data showed some mention of protest prior to the Cairo embassy attack, but none in the Benghazi consulate attack. The results suggest that the Cairo embassy attack was a spontaneous uprising, whereas the Benghazi attack was planned.

This example shows how network analytics and language technology can be combined for rapid forensic assessments in the immediate aftermath of the event. Key challenges include using sources from multiple languages (including recognized languages and variants used in social media tools, such as replacing certain sounds with numbers), reconciling differences in data in English and non-English press, identifying the key stakeholder groups and the relations among them, and merging data from diverse sources in near real time.



Expectation	Find
Tunisia first event	Yes and spike in violence
Libya and Egypt High violence, revolution, anger against revolutionary activity, overthrow	Yes
Egypt Continued violence and anger against protests	Yes
Tunisia, Yemen Less violence, less revolutionary activity, overthrow	Yes – Yemen weak on revolution
Syria Increasing violence	Yes
Qatar, UAE Low violence, low revolutionary activity, no state change	Yes
Kuwait, Oman, Morocco Low violence, moderate revolutionary activity, minor state change	Yes – Oman strongest, Morocco slow
Iraq, Iran High violence, low revolutionary activity	Yes

FIGURE 2.5 Comparison of dynamic network analysis prediction in Egypt (right) with actual events associated with the Arab Spring uprisings (left). (Right) Mean change in the revolution and violence beliefs for each country and each month. Trend lines represents the change in belief that revolutionary activity is good (gray line) or that violence is needed (black line). Vertical lines represent the intercountry ranges (black lines) and the intracountry ranges (gray lines). Large black dots indicate the month (October 2011) when the model first forecasts that a revolution is likely to occur. The revolution actually began in November 2011. Peaks indicate a forecasted escalation in that activity. A major protest in Egypt occurred in January 2011. Revolutionary activity was indicated when there was increasing sentiment that nothing was being done to stop revolution (negative revolutionary activity beliefs), and the violence belief was noticeably positive. SOURCE: Data from Joseph et al. (2014). NOTE: UAE = United Arab Emirates.

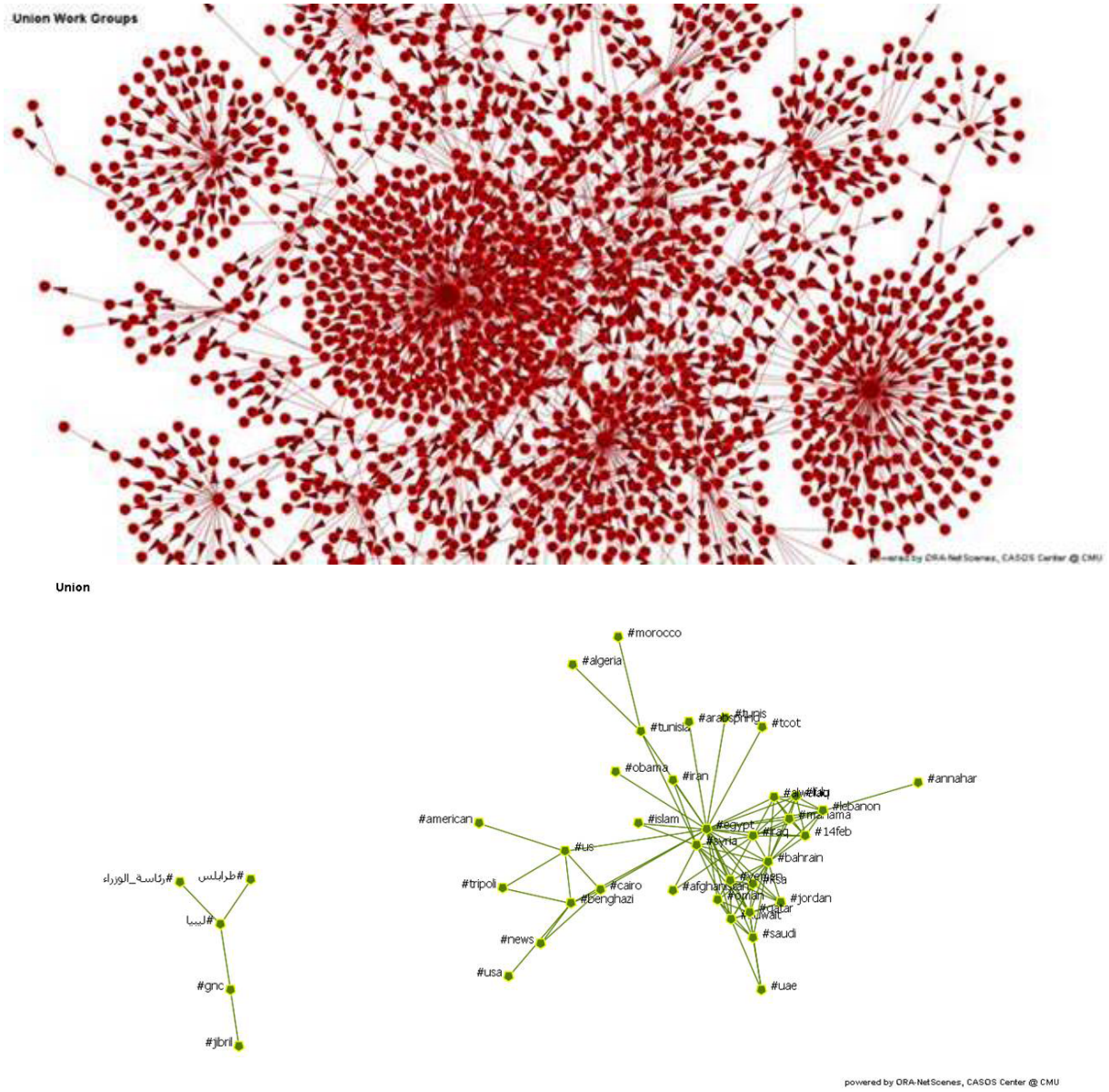


FIGURE 2.6 Illustrative Twitter networks from the September 2012 Benghazi consulate attack. (Top) Overall tweet network, showing that tweets from a few sources are retweeted by others. The Tweeter IDs have been removed and each arrow indicates that a tweet by the origin was retweeted by the destination node. This leads to the typical twitter starburst pattern. (Bottom) Hashtag network. Each node is a hashtag from the tweets, tweeted by those shown at the top. A link means that both hashtags were used in one or more tweets. Note that there is no direct linkage between the Arabic topic group (left) and the English topic group (right). SOURCE: Carley et al. (2014). With permission of Springer.

Tactical Question: What are the relations among the major ethnic and political groups surrounding the Islamic State in Iraq and Syria (ISIS)?

To determine how to respond to ISIS, it is important to know who are the key groups or actors, how they are related, the basis for their power, and which group can be expected to influence which other group. One might begin by developing a list of all relevant groups—terror groups, militias, political groups, and ethnic groups—and then using network models to identify the relations among the groups. With geospatial intelligence data, such an assessment can provide insight into which groups are highly relevant in which countries. For example, Figure 2.7 shows the prevalence of discussions about ISIS per country over time. Discussion about a group is generally a reasonable indicator of their presence in a region.

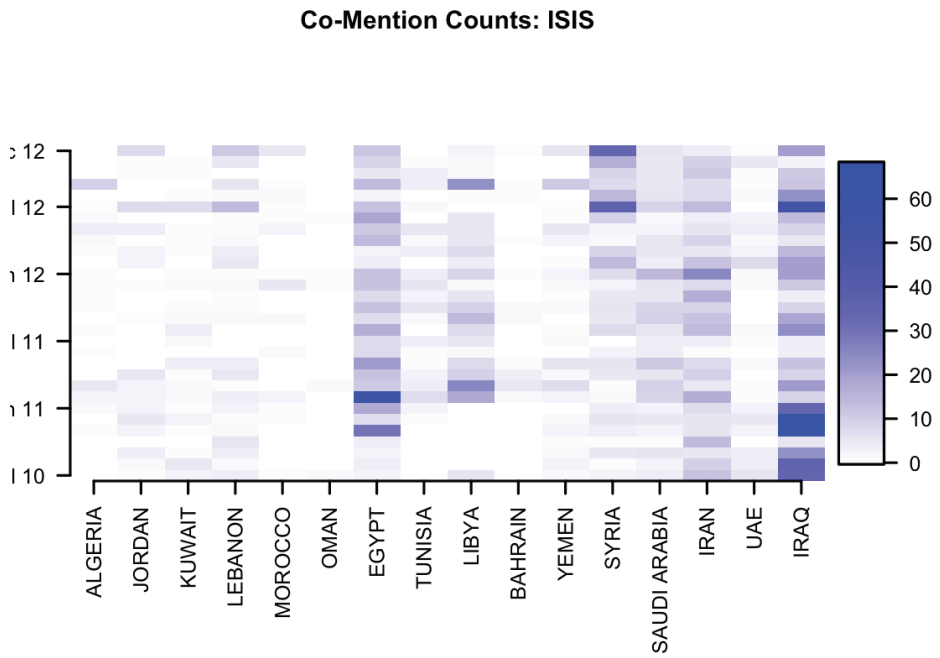


FIGURE 2.7 Presence of ISIS in news media by country from July 2010 to December 2012. A cell indicates the volume of news articles discussing ISIS and that country in that month. The darker the color the higher the volume of discussion. The figure shows that ISIS was in the news and part of the discussion about tensions in the Middle East in the majority of these countries before the start of, and during, the Arab Spring.
NOTE: UAE = United Arab Emirates.

Network models can also provide insight on influence and power. For example, the copresence of particular groups in articles is often used as a measure of relations among those groups. Figure 2.8 is a network diagram illustrating alliances and hostilities of some key groups to ISIS, based on 75 online news articles and open-source reports discussing ISIS and its relation to various groups between January and October 2015. The diagram shows that ISIS has built alliances with a large number of groups, whereas the groups opposed to ISIS are less centrally coordinated and may be hostile or indifferent to one another.

Key challenges in such an analysis include setting the strength of and confidence in the links, understanding the basis for links and predicting changes in them (e.g., identifying and visualizing the underlying sociocognitive map), assessing regional differences in groups (e.g., Sunni groups have different concerns and different levels of commitment for or against ISIS and the Iraqi military, based on their level of religiosity and location), and identifying new groups as they emerge.

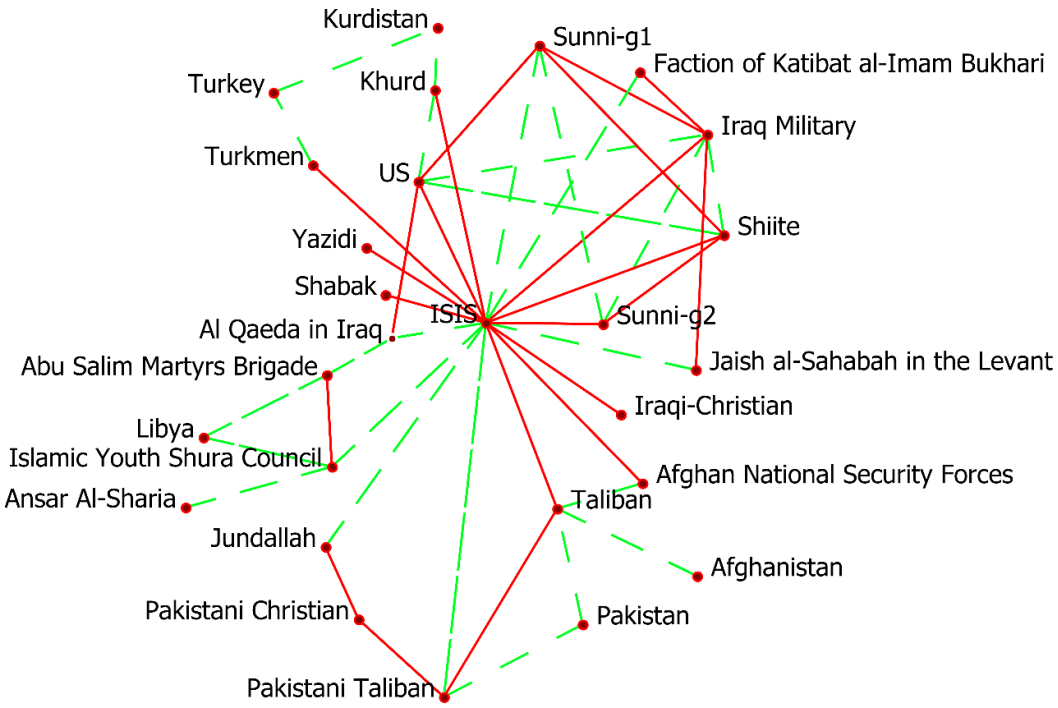


FIGURE 2.8 Alliances (green lines) and hostilities (red lines) between selected groups and ISIS. This image is illustrative and does not represent the complete set of groups and their relations associated with ISIS in 2015. It is based on network data extracted from approximately 75 articles from LexisNexis, coded for groups and mention of an alliance or an act of hostility between these groups.

EXAMPLE 5: FOOD AND WATER SCARCITY

A growing population and increasing income, especially in the developing world, is increasing global demand for food, water, and energy. Agricultural demand, for example, is projected to increase 60 percent by 2050, relative to 2005 levels (FAO, 2012). These trends—along with the interdependent nature of food, energy, and water, as well as their sensitivity to external factors such as climate change—pose significant challenges for managing critical resources. For example, increasing the production of biofuels increases water demand and shifts water resources away from food production. Because decreases in food and water have the potential to exacerbate humanitarian crises (Clapper, 2015), understanding the future evolution of these systems and their interactions is critical.

Strategic Question: What regions are vulnerable to food and water shortages in the future due to interactions between food, energy, water, and climate?

Traditionally, energy, food, and water systems have been planned, developed, and analyzed independently. However, integrated assessment models, which couple representations of physical and social systems, are beginning to be used to understand some of the interconnections and interdependencies between water and other systems and to project future outcomes for food, energy, and water. For example, three integrated assessment models have recently been expanded to include water, enabling the impact of changes in population, income, technology, and climate on the supplies and demands for water, energy, and food to be assessed. Such models are used by a variety of government agencies (e.g., the U.S. Department of Energy and the U.S. Environmental Protection Agency) and

nongovernment organizations (e.g., the World Bank and ExxonMobil) to analyze the effect of technology, policy, and climate on economic systems.

Analyses using these models have quantified the expected amount and distribution of future water scarcity around the world under different scenarios (Hanasaki et al., 2013; Hejazi et al., 2014; Schlosser et al., 2014). These studies calculated scarcity as the ratio of water desired (i.e., demand if water were unlimited and free) to the amount of water that is actually available. Water desired was determined using an integrated assessment model, which accounts for future changes in population, income, energy demand, agricultural demand, and other factors. Water supply was determined using hydrology models (e.g., monthly water balance models), which simulate runoff and streamflow under different climatic conditions. The results suggest that increases in water demand and changes in water availability in regions such as India and China may lead to increased water scarcity in those regions (e.g., Figure 2.9).

The same models were used by Kim et al. (2016) to examine the effect of limited water availability on agricultural systems, but in this case, water prices were adjusted to prevent the demand for water from exceeding the supply in any region. Thus, higher water prices in a particular region would indicate higher water scarcity, and cause reductions in the consumption of water-intensive goods and shifts in agriculture production. Figure 2.10 illustrates future shifts in the production of corn, rice, and wheat under three scenarios: lowest-cost groundwater, highest-cost groundwater, and baseline median estimates for groundwater cost. The figure shows that increased water scarcity forces the greatest shifts in the regional production of food, and that reductions in crop production in a water-scarce region are offset by increased crop production in regions with more abundant water.

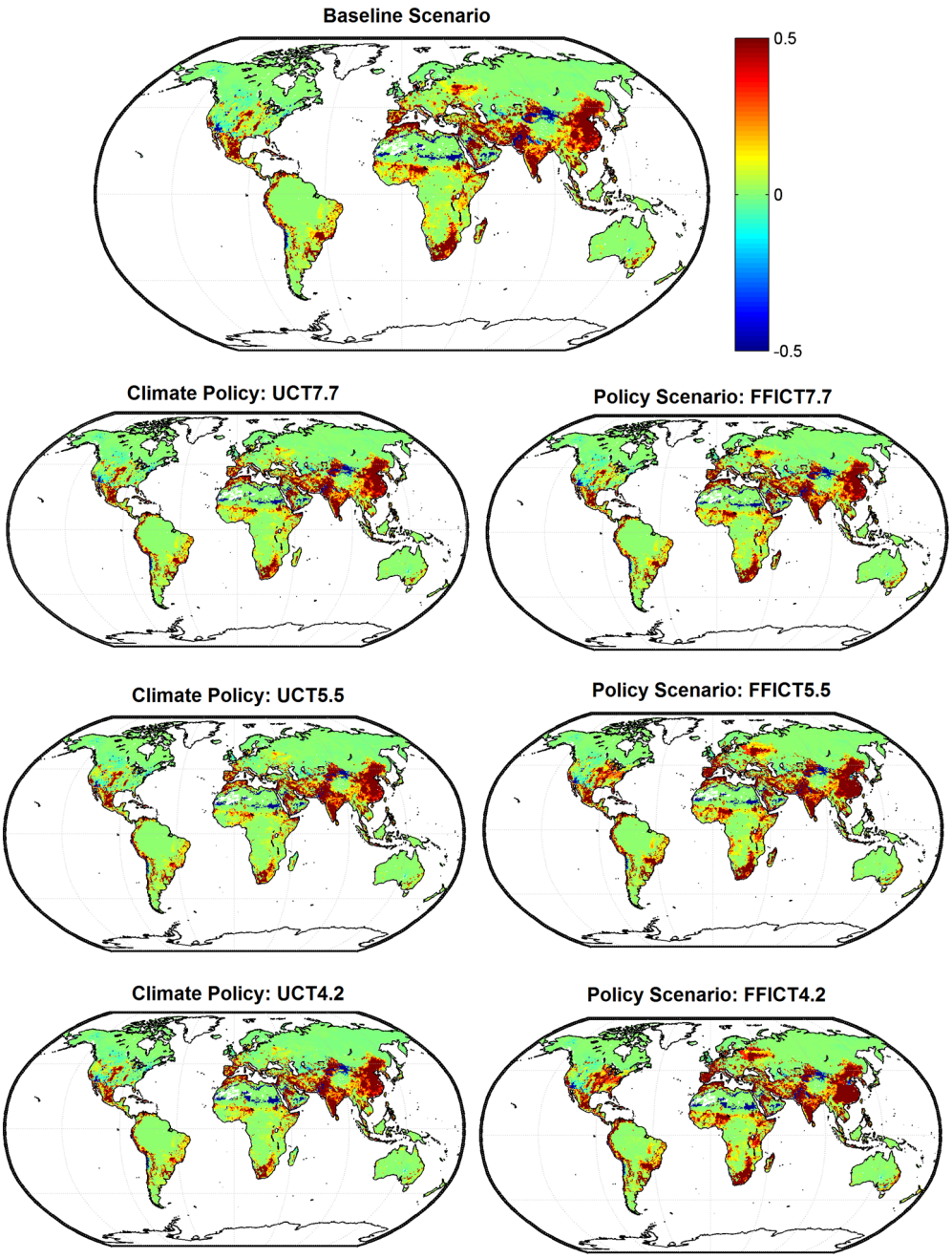


FIGURE 2.9 Changes in water scarcity from 2005 to 2095 under different climate change mitigation policies, including no climate policy (baseline), a carbon tax regime covering all sources of carbon emission (UCT), and a carbon tax regime covering only fossil fuel and industrial emissions (FFICT). Red indicates severe water scarcity conditions in 2095 relative to 2005. SOURCE: Hejazi et al. (2014).

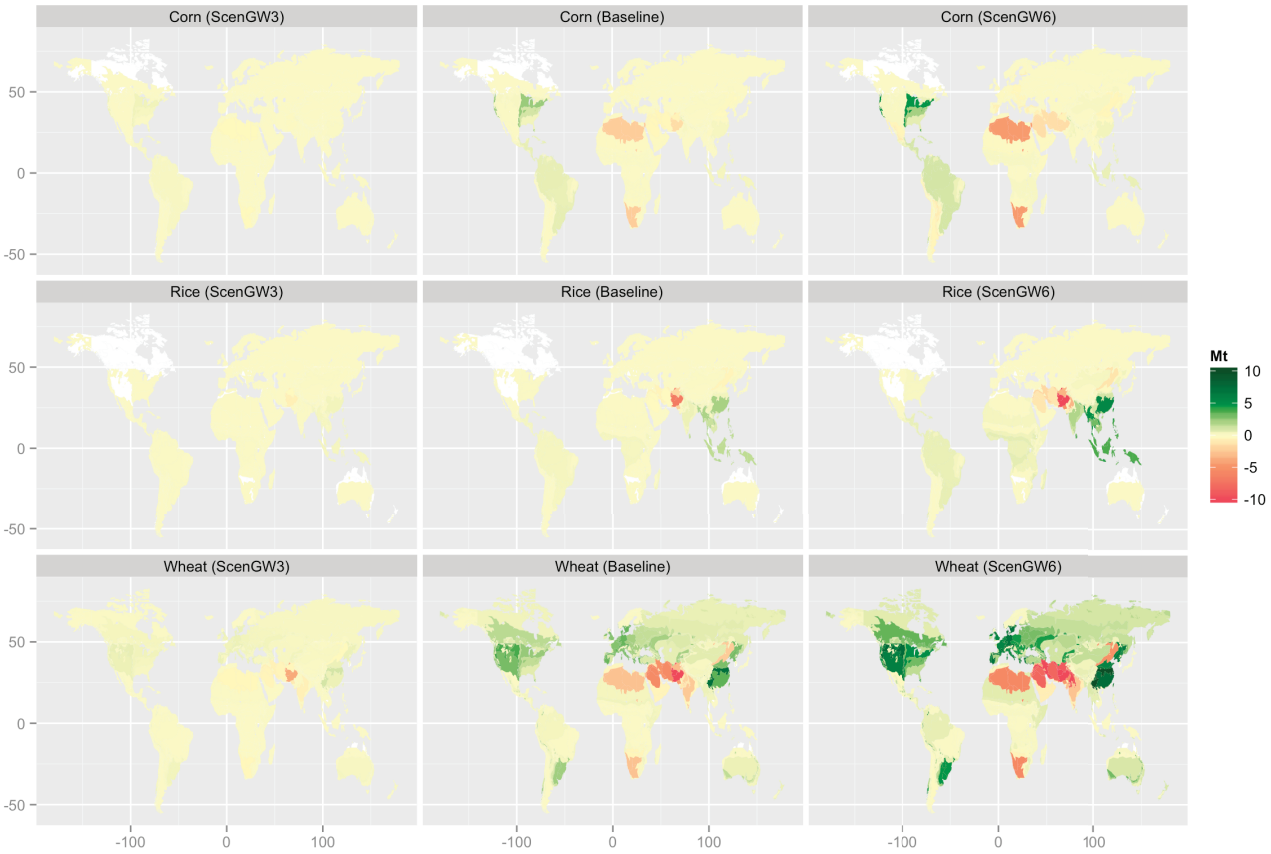


FIGURE 2.10 Changes in production of corn, rice, and wheat (in million tons [Mt]) in 2100 for three water price scenarios compared with an unconstrained water scenario. (Left) Lowest-cost groundwater scenario (ScenGW3). (Middle) Baseline scenario, using median estimates of groundwater cost. (Right) Highest-cost groundwater scenario (ScenGW6). SOURCE: Kim et al. (2016). With permission of Springer.

3

Model Foundations

Models do not stand alone, but rather exist in an environment that includes the data available, the computational and data infrastructure, methods for analysis and model assessment, and tools for interpretation, visualization, and understanding. In addition, the system needs people skilled in choosing, tailoring, and running the models and interpreting their output. Together, these elements provide a basis for making inferences about real-world systems (e.g., Figure 3.1). This chapter provides an overview of these elements, using key issues and questions to illustrate how the National Geospatial-Intelligence Agency (NGA) could think about them from a practical standpoint. The first section focuses on models, including whether to develop new models or use existing models and what types of models to use in the investigation. The second section focuses on data, including sources and methods for handling them in the analysis. The third section covers assessment of a model's connection to the real-world system, with particular emphasis on verification, validation, and uncertainty quantification. The fourth section examines what types of computation (e.g., high performance, data intensive, and spatial) are necessary for an investigation. The fifth section includes a discussion of tradeoffs among these elements, which influence the type of information provided, the time or cost required to provide it, and the ability of end users to analyze and understand the results. The chapter ends with a brief summary and conclusions.

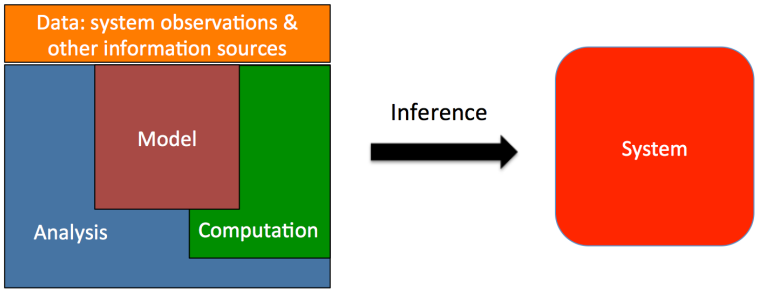


FIGURE 3.1 A model-based investigation to better understand or predict a complex, real-world system will require some combination of model, data, analysis, and computation. The process is typically iterative, using initial results and inferences as well as human judgment to improve data collection, modeling, computation, and/or analysis methods.

MODELS

The complexity of real-world physical–social systems makes it virtually impossible for NGA to build and maintain all the models that would be useful for producing geospatial intelligence. Consequently, an overarching question to be addressed before embarking on any modeling approach is the following:

Can "off-the-shelf" models be used in the investigation, or do new models have to be developed?

Developing a model from scratch makes it possible to produce output that directly addresses the goals of a model-based investigation. Relatively simple models can be easier to build than to adapt from existing models. In contrast, development of complex models is often a resource-intensive, experts-only task, requiring design, coding, debugging, testing, verification, validation, and documentation. If these tasks are not done thoroughly, the resulting model or software might be usable only by the developers and may become fragile or unusable when those individuals move on to other assignments. Using existing model output or previously developed models can greatly reduce the costs of the investigation. Previously developed models will also likely have undergone some quality checking, verification, and comparison to actual system data. However, because they were developed for a different purpose, existing models or model output will likely need some adaptation or augmentation for the investigation at hand. Advantages and disadvantages of developing models, combining existing or new models, using existing model output, and running an existing model are outlined below.

Developing models. If appropriate system models are not available, they will have to be developed. Model development takes into account the purpose of the model, the system to be explored, computational and data constraints on the investigation, and the acceptability of the various approximations available. Model purposes include exploring the range of possible outcomes, making predictions, developing theory, explaining phenomena, and filling in missing data. Choices about the model development include the following:

- Which phenomena and behaviors need to be captured in the model, recognizing that it is not necessary or feasible to model everything;
- The granularity, resolution, or fidelity needed to properly represent (for the model's purposes) each of those phenomena or behaviors, and whether this varies with space and/or time;
- Issues of representation, such as spatial and temporal dependencies (e.g., autocorrelation and heterogeneity) that need to be captured, whether the model should be deterministic or stochastic, and whether it should be steady state or evolving according to physical, social, or empirical rules;
- What aspects of the model should be based on empirical data, which should be based on theory or rules, and what level of empiricism is acceptable; and
- How the submodels will be combined, including what types of coupling and feedbacks are necessary.

Models may also be developed within a preestablished modeling framework, which avoids having to create the full infrastructure and all metrics from scratch. With this approach, the model developer first selects a framework that has the features needed for visualization, metrics, and data handling (e.g., Stella or i-Think for system dynamics, NetLogo or Repast for an agent-based model, and ORA or R for a network model). Then the model is built in the framework, the analysis is carried out, and the results are presented as images, maps, diagrams, or other visual depictions. In general, modeling frameworks that are interoperable with many other tools and that have a wide range of allowable input–output formats are preferred.

NGA deals with systems that evolve in time and space, and it also needs to make sense of large volumes of empirical intelligence data. Example questions associated with some of these issues are discussed below.

Does the investigation require analysis methods to account for spatial, space-time, or network dependence structure?

Perhaps all models and data of interest to NGA have a spatial, space-time, or group or social network dependence structure. Examples include geospatial data, social media data, specialized spatial data or geo network models, or output from a social system model evolving in a spatial environment. If the system exhibits spatial or network dependence, standard statistical approaches that rely on independence of the observations and entities will not be appropriate for analysis. Rather, specialized methods are used for spatial analysis (Shekhar et al., 2011), machine learning (Bishop, 2006), and network analysis (Kolaczyk and Csárdi, 2014; Scott, 2013; Wassermand and Faust, 1994). Use of these methods requires computational approaches that can deal with space-time data and systems, huge databases and intricate calculations for estimating network properties, or specialized metrics and algorithms, such as those developed for spatial network data.

Are empirical models, process models, or some combination needed for the investigation?

Empirical models combine system data with simple mathematical models to lend insight about the system. A simple empirical model relating the heights of parents to those of their offspring is shown in Figure 3.2 (left). A more complicated dynamic empirical model of sea-surface temperatures, which produced a reasonably accurate forecast of the 1997 El Niño event, is shown in Figure 3.2 (right). Empirical models are often used for describing "normal" behavior, which is helpful for detecting anomalies. The Legionnaires' disease example in Chapter 2 used empirical concepts to help locate the geographic source of the outbreak.

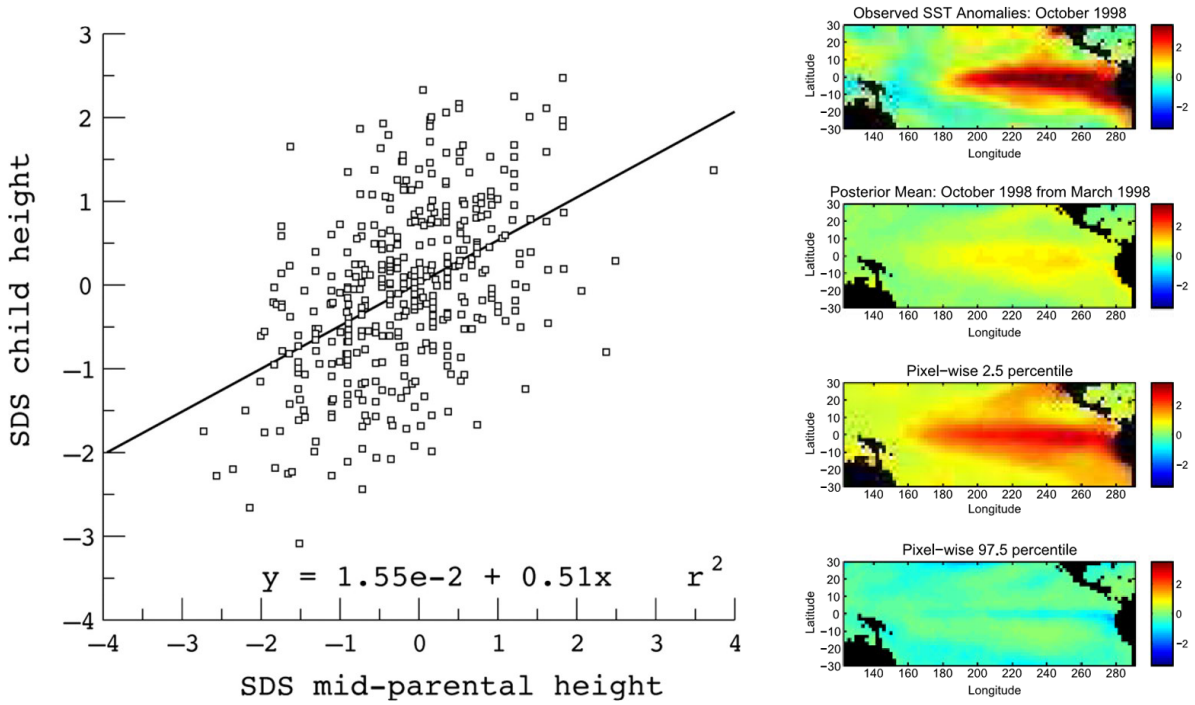


FIGURE 3.2 Empirical models may be static and simple or dynamic and complicated. (Left) Correlation diagram showing the empirical relationship between parents' height standard deviation scores (SDS) and those of their offspring. SOURCE: Reproduced from Wright and Cheetham (1999), with permission from BMJ Publishing Group Ltd. (Right) Remote sensing observations and predictions from a dynamic empirical model of sea-surface temperature (SST) evolution. SOURCE: Wikle and Hooten (2010).

In contrast to empirical models that do not attempt to represent the inner workings of a system, process models are typically built to simulate the processes and behavior of physical or social systems, which often evolve over time. Such models, which can be based on theory (e.g., equations of fluid flow) and/or rough rules (e.g., presumed response of individuals to a situation for which no data exist), examine the possible responses of a system to changes in conditions. Components can be coupled to allow two-way interaction, and interactions that determine the response of a system to a disturbance (feedbacks) should be represented. Developing process models is typically demanding and involved. If many components and feedbacks are required to represent the system, model development will require substantial labor, time, and computing resources.

Model run time is also an important consideration, because model-based analyses can require anywhere from tens to millions of model runs or be so complex that days to months are required to complete the simulation. Run time can be reduced by increasing computational resources and efficiency or by developing reduced models. Reduced models use coarser, simpler, or fewer representations of processes than a full model. They could be based on simpler mathematical representations (e.g., reduced-order models; Willcox and Peraire, 2002), a few processes (e.g., motivated metamodels; Davis and Bigelow, 2003), or a response surface trained on an ensemble of full model runs (e.g., emulators; Sacks et al., 1989). Developing a reduced model that captures only the most important features for the application at hand may prove advantageous for analyses that require many model runs or long simulation times.

Most computational models are not based solely on first principles; rather, they contain many empirical specifications and parameters that help define the system being studied. As understanding increases, the number of components and feedbacks that are represented in the model, rather than specified empirically, tends to increase (e.g., Figure 3.3). Such complex models strike a balance between retaining process fidelity (consistency of subsystem models with observations of process-specific variables) and system calibration (consistency of the system with observations of system-level variables). Complex model systems can be relatively brittle and challenging to construct, but the effort is warranted by the confidence that the calibration does not drive the system out of consistency with observations at the process level. In other words, the model will not produce the right answer for the wrong reasons, simply by tuning.

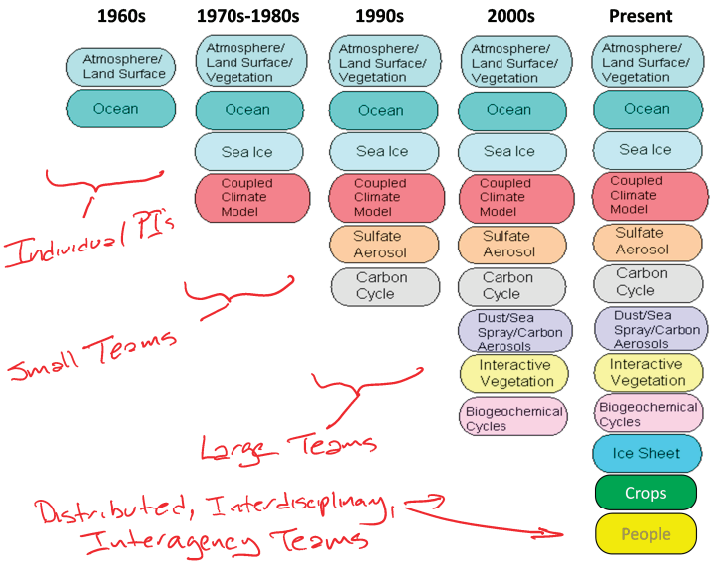


FIGURE 3.3 Evolution of climate model development. Over the years, the number of components being modeled has increased, as has the size and disciplinary expertise of the teams developing the models. SOURCE: Modified from Washington et al. (2009).

Combining models. Models built for different purposes, using different techniques, or simulating different types of subsystems are often combined into a larger model of a real-world system. The ability to use diverse models together is referred to interoperability, and the selection and combination of relevant models to generate meaningful results is referred to as composability. Combining multiple models allows more aspects of the system to be modeled and explored, supporting the formulation of additional conclusions, but it also requires more teamwork, because a team of model developers is often needed for each submodel (e.g., Figure 3.3). Having an arsenal of subsystem models (e.g., representing different processes and containing different levels of detail or resolution) that could be combined for different purposes could make it easier for NGA to rapidly provide information on emerging threats. Consequently, a key question is the following:

How should multiple models be combined for the investigation?

Combining multiple models to represent a complex system is challenging because the behavior of a complex system cannot be expressed as a sum of the behaviors of the components or subsystems; complex interactions and feedbacks between processes can dominate system evolution. Effectively representing the information that must travel between subsystem models to establish feedbacks and interactions also poses semantic and software challenges. For example, common, compatible representation and language are needed to correctly align dynamics and express information in coupled domains. In addition, most models are not designed to be used in plug-and-play architectures, and standards and protocols for interoperability and composability (covering technology, syntax, semantics, etc.) need to be developed (Davis and Andersen, 2004; Hofmann, 2004). Finally, significant software challenges arise from building, executing, and analyzing a complex model—often containing submodels with different spatial and temporal domains, different resolutions, and different data sources—in a distributed computing environment. Combining these models often involves a data fusion exercise.

Obtaining useful results from combined models depends on appropriately capturing the interactions among the different subsystems, which may be used in parallel, as inputs of one subsystem for another, or as joint contributors to a third (see Appendix A). Figure 3.4 illustrates the linkages among diverse submodels that must be properly represented in a comprehensive model of the human–climate system. A key consideration is how strongly the subsystem models being combined are expected to couple and interact with one another. The pirate example in Chapter 2 illustrates one-way coupling, because outputs of the weather and wave model serve as input to the pirate behavior model, but the pirate activity does not feed back to the weather model. In the housing bubble example, households, banks, and their economic environment all interact with one another as the system evolves.

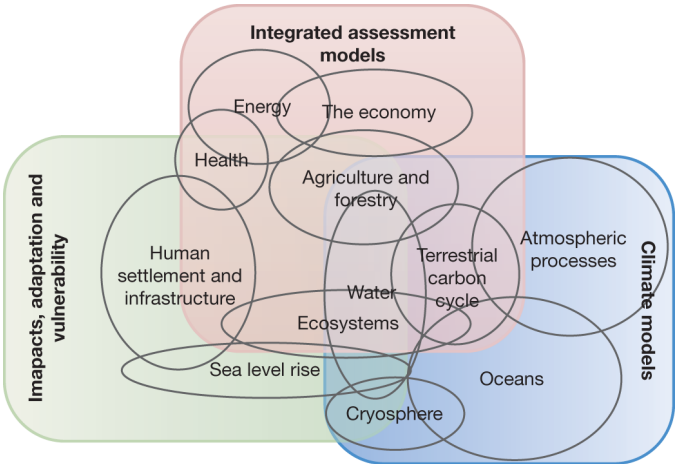


FIGURE 3.4 Linkages of components in comprehensive models of the human–climate system. SOURCE: Moss et al. (2010). Reprinted by permission from Macmillan Publishers Ltd.: Nature.

Using existing model output. Perhaps the simplest way to exploit a model is to use output from previous model runs. For example, existing weather and wave forecasts were used to help predict the likelihood of pirate activity (see Chapter 2). A key disadvantage of using existing model output is the lack of control and the lack of knowledge about what went into the model, especially when documentation about the quality assurance and assessment procedures is sparse. In some cases, the model developers may be available to discuss these procedures as well as the pros and cons of using their model output for the investigation at hand. They may even be willing to carry out additional runs to support the investigation.

Analyses of model output from a coherent ensemble—such as a collection of model runs from different models that simulate the same system, or a collection of model runs that systematically vary model inputs or parameters—can provide a fuller picture of the behavior of that system. And when multiple models provide similar results, confidence in the results increases. Knowing that an ensemble of model runs will be required will affect how a model is selected or developed for the investigation. Consequently, a relevant question is the following:

Will model-based analyses require ensembles of model runs?

A number of climate investigations make use of model output collected in the Coupled Model Intercomparison Project repository.[1] For example, Tebaldi and Knutti (2007) and Smith et al. (2009) used the output to develop hierarchical models to project future climate behavior and to quantify uncertainty in the projections. Such approaches have also found success in weather forecasting (Raftery et al., 2005). Precomputed ensembles of model runs are also used for model calibration (Higdon et al., 2008; Kennedy and O'Hagan, 2001), sensitivity analysis (Helton, 1993; Saltelli et al., 2008), uncertainty propagation and analysis (Oakley and O'Hagan, 2002; Spanos and Ghanem, 1989), and approximate Bayesian computation (Beaumont, 2010; Fearnhead and Prangle, 2012). For example, both the pirate and the housing bubble examples in Chapter 2 make use of sensitivity analyses produced from ensembles of model runs. The availability of computational resources will affect how such ensemble-based analyses are carried out.

When models are empirically driven, as in the Middle East example in Chapter 2, an ensemble can be created using bootstrapping and sampling techniques. For simulation models, these ensembles are created by varying the parameters through a distribution and the results can be thought of as showing the range of what is feasible. For more empirical models, such as network models, the ensembles are created by sampling from existing data and can be thought of as showing the robustness of the metrics and model result to missing data.

Setting up and running an existing model. When a model is useful, but the existing output is not appropriate for the investigation, a new computational run or ensemble of model runs can be specified and carried out. An input file (or files) typically specifies initial and boundary conditions, along with key model parameters. Once initialized, the model can be run, producing a file (or files) containing the model's output. Scripting languages, such as python or R, can be used to carry out runs over different input configurations, producing output for sensitivity studies or other analysis. The DAKOTA software from Sandia National Laboratory (Eldred et al., 2007) has utilities for managing many model runs on a variety of computing systems, as well as a collection of analysis capabilities to process the resulting model output.

Existing models can be run using an iterative approach, in which the results of the model run (or ensemble of model runs) are used to determine a new input setting (or settings) for the next model run. Examples of analysis methods that make use of this iterative running of models include the following:

[1] See http://cmip.pcmdi.llnl.gov.

- Analysis methods for inverse problems (Kaipio and Somersalo, 2006; Tarantola, 2005),
- Model calibration and optimization approaches (Eldred et al., 2007), and
- Data assimilation approaches (Courtier et al., 1994; Evensen, 2009).

Iterative methods have proven effective for successive computational model runs that can be carried out relatively quickly (e.g., fractions of a second, or minutes). If the investigation requires a complex model with a large set of inputs, even high-performance computing resources may not be sufficient to make this approach feasible. Strategies to speed the required iterations include linearization (e.g., Kaipio and Somersalo, 2006), use of reduced models, and use of adjoint or derivative information to facilitate the parameter search (Butler and Estep, 2013; Martin et al., 2012).

DATA

Data—measurements or observations of the real-world system—link the model to the real-world system it is intended to represent. Nearly all data of interest to NGA have a geographic or locational component (see Figure 3.5). However, the sheer variety of these data—collected at different scales and for different purposes and stored in different formats—poses challenges for finding relevant, high-quality data, integrating them into a model-based investigation, and accounting for their nature in the analysis. Example questions concerning data sources and methods for ingesting data into the analysis are discussed below.

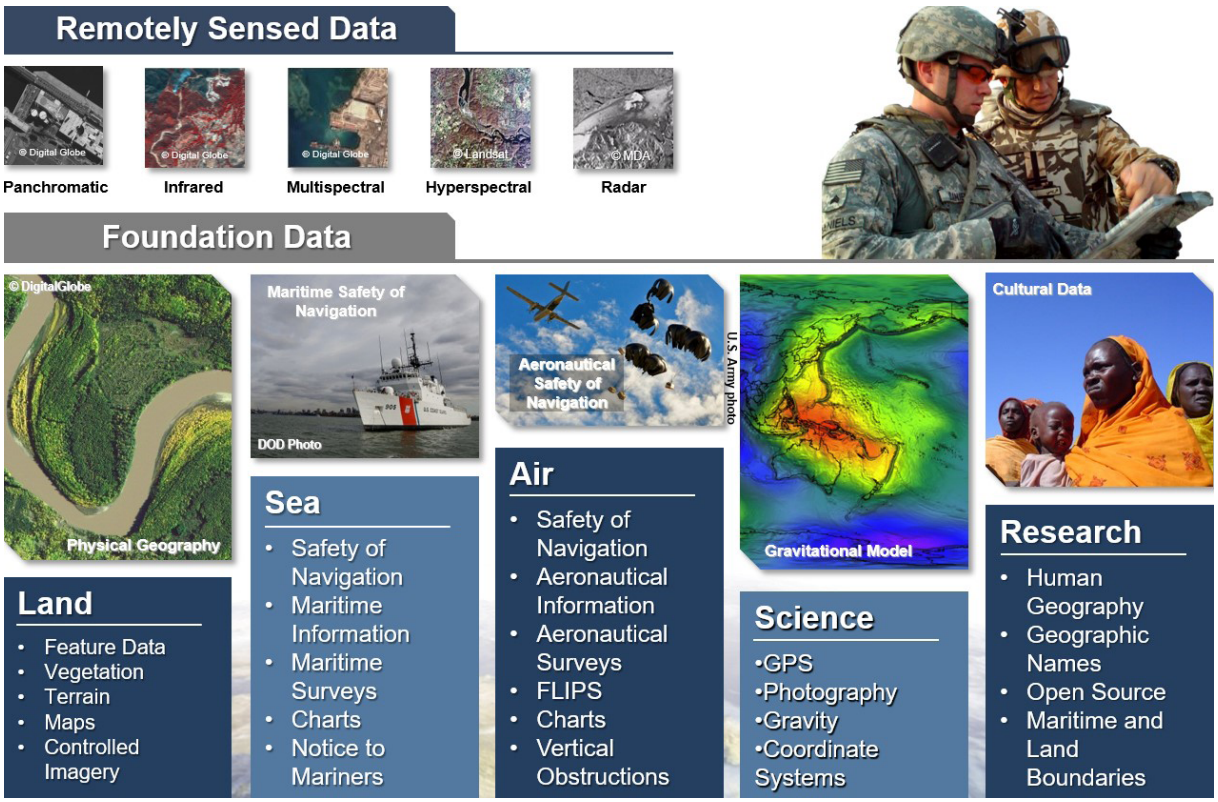


FIGURE 3.5 Examples of the wide range of data used by NGA. SOURCE: NGA.

What data are available for the investigation?

NGA's geospatial intelligence mission is focused on places and activities outside of the United States. A wealth of environmental data is readily available from satellites, scientific sampling campaigns (e.g., Argo ocean floats), and intergovernmental agreements (e.g., meteorological and hydrologic data from the World Meteorological Organization), particularly at global and regional scales. Higher-resolution environmental data as well as data for areas with limited scientific infrastructure (e.g., Africa) are comparatively sparse. Relevant social system data are available from experiments (e.g., natural experiments for observing social systems under different conditions [Rutter, 2007]), administrative and economic records, newspapers, and social media. For social media data, it may be difficult to determine what data are available as well as what data can be extracted and by whom. For many types of data, access is restricted due to privacy, confidentiality, competitive advantage, or national security issues.

It is important to understand how the data were collected, their potential sources of error and bias, and their quality, especially for data that are directly relevant to the key questions of interest. For remotely sensed data and imagery, which are central to NGA investigations, it is necessary to account for positional accuracy, blurring properties, registration errors, and spatiotemporal resolution to properly integrate the data into a model-based analysis. In addition, data products derived from these sources (e.g., land cover classification) will have their own uncertainties that may need to be accounted for in subsequent analyses.

When data collected from diverse sources are used, care has to be taken to account for the temporal, spatial, or sociocultural contexts of the data so that the data can be used collectively in a model. Particular data sources may have biases, errors, or uncertainties that are not known a priori. Tailored models that account for the possibility of systematic errors, biases, and outliers in some data sources may be needed to better represent the uncertainty. This will reduce the chance that otherwise plausible conclusions are discarded because they are not consistent with an unreliable information source.

The quality and accuracy of the available data can also affect how the model-based investigation is scoped as well as the robustness of the conclusions. If high-quality, directly relevant data are available, it may be preferable to use simpler, data-driven models to answer key questions. The use of directly relevant, but diverse, data can also lead to more robust conclusions; and capturing that diversity may require more detailed modeling and analyses. If only indirect system data are available, inverse models or other model-based analysis approaches are required to make inferences about features of interest in the real-world system. Special computing infrastructure and methods may be required to search through various sources of data and associated metadata, determine what data may be useful for a particular investigation as well as their nature and pedigree, and store particularly relevant data and analyses for future investigations.

How will data be ingested in the model-based analysis?

The nature of the data drives how it is handled in a model-based investigation. Data may be massive, requiring special cyberinfrastructure (cloud, high-performance, or high-throughput computing) to process; they may be streamed at a high rate, requiring specialized analysis approaches to carry out estimation tasks with a single pass through the data; they might accrue over time (e.g., hourly or daily), requiring data assimilation; they might be observations from a single experiment, requiring an inverse analysis; and they might not exist at all, requiring expert judgment to inform the models. Often modeling approaches, analysis methods, and computational infrastructure need to be developed in concert to efficiently leverage the data being collected for the investigation.

Ingesting geospatial information into models raises special considerations. For example, many physical process models, such as general circulation models, use a spherical reference frame to specify locations on the surface of Earth. However, geospatial intelligence applications require more accurate representations of Earth, such as ellipsoid representations and nonparametric representations that use land-based geodetic reference points for localization. In addition, general circulation models often use raster-type data to organize geospatial information

as a matrix of cells (e.g., in a latitude–longitude grid), each of which contains an information value such as temperature. In contrast, geospatial information comes in many forms, including raster (grid), vector (e.g., objects such as points, lines, and polygons), and network (e.g., a collection of nodes and edges to represent a mathematical graph). For example, a map of water bodies in vector format may use points to represent locations of wells, line strings to represent center lines of rivers, and polygons to represent the footprint of lakes. An urban street map in graph format may include nodes to represent road intersections and edges to represent the road segments connecting adjacent intersections.

Data often require some form of preprocessing to be useful to a modeling effort. For example, the Middle East example in Chapter 2 used processed text data from newspapers, Twitter, and other media to determine linkages among social and political entities. Most sensors record electronic quantities, such as voltage, that need to be processed and modeled to turn them into properties reflective of the real-world system, such as temperature. In climate and weather modeling, it is common to produce data products, which combine raw measurements with modeling to yield something close to raw observations, but interpolated over a regular, spatial grid. Figure 3.2 (right) is an example of such a data product.

A number of methods can be used to ingest data into the model. Perhaps the most basic approach is to "pre-ingest" the data within the model itself, producing model results that have been constrained to match known databases. The model has to be set up appropriately for the investigation (e.g., specify initial and boundary conditions), but a formal model calibration and uncertainty assessment may not be required. An example of this approach is the MCNP-X model (Hendricks et al., 2008), which simulates the transport of nuclear particles through specified materials. The code makes use of a large, vetted database of nuclear cross-section experiments, and so it is not necessary to determine scattering properties of various nuclear particles in different media.

More commonly, some form of analysis is used to combine a model with system observations (and perhaps expert judgment). In cases where the observations are fixed (e.g., static, historical measurements), model calibration approaches (Kennedy and O'Hagan, 2001) and inverse methods (Tarantola, 2005) can be used to estimate uncertain model parameters or uncertain initial or boundary conditions. Uncertainty quantification tools such as model emulation (Conti et al., 2009; Sacks et al., 1989) and sensitivity analyses are often useful in such exercises. In cases where the model needs to be updated repeatedly to produce time-sensitive predictions, data assimilation techniques are used to infuse data into the model. U.S. numerical weather models, which produce weather forecasts every 6 hours, are a classic example.

MODEL ASSESSMENT

In any model-based investigation, it is crucial to understand the strengths and weaknesses in the model's connection to the real-world system. The processes of verification and validation have a long history in supporting the broader task of model assessment, particularly in the engineering and physical sciences (NRC, 2007b, 2012; Oberkampf and Roy, 2010; Oberkampf et al., 2004; Pace, 2004). Verification assesses the adequacy of the computational model's fidelity to the mathematical model, and validation assesses the adequacy of the computational model's representation of the real-world system. When data are not available for quantitative comparison, assessment may include comparing different modeling approaches or qualitatively comparing model results with process understanding to map out what features of the model are trustworthy and how the model and reality will likely differ. Uncertainty quantification also supports model assessment by estimating uncertainties in model-based predictions (e.g., Smith, 2013). Example questions concerning these model assessment tasks are discussed below.

How will verification, validation, and uncertainty quantification be carried out to support model assessment?

How model assessment tasks are carried out depends on the properties of the model, the availability of relevant data, and the nature of the key questions in the investigation. Verification can be a demanding and involved process

for large-scale computational models of physical processes, which use iterative algorithmic schemes to solve large systems of differential equations (Oberkampf and Roy, 2010; Roache, 2002). In contrast, verification is less of an issue for empirical models, because the mathematical representations are relatively simple and tested in software. Validation of empirical models focuses on how representative the data used to estimate parameters are for the key questions of the investigation. Here, standard model checking approaches from the statistical and machine learning literature are relevant, such as holding aside some data for final checking. Uncertainty quantification is typically built into the models (e.g., statistical models commonly produce prediction and parameter uncertainties). Models can also be built to facilitate assessment, in particular by constructing them to export data into the same analysis tools used for the relevant real-world data. Many modeling frameworks have such built-in analysis toolkits.

In extrapolative (i.e., new, outside of previously tested or observed conditions) settings, validation and uncertainty quantification can be challenging; this holds for speculative models of social systems as well as for process-based models based on well-understood theory. This is because common approaches assume some form of process stationarity (i.e., tomorrow will be like yesterday) and continuity (the system changes slowly in time and space) to estimate model errors and quantify prediction uncertainties. When the real-world system has evolved to a new regime, making present behavior unlike past behavior, these approaches for validation and uncertainty quantification are not appropriate. Examples of assessments of agent-based models appear in the work of North and Macal (2007).

In dynamic environments, the model needs to be updated repeatedly with new data. The methods required to assess these models are often relatively intricate, requiring approaches to update estimates and uncertainties, often in real time, and often involving large volumes (e.g., gigabytes to terabytes) of data. However, the repeated arrival of new data presents an opportunity to directly compare model predictions to reality.

Models built for different purposes or using different frameworks are often used in a multimodeling framework as a form of validation and/or confidence building. In particular, if all models suggest the same outcome, then one can have greater confidence in it. Using multiple models from diverse theoretical traditions can provide a more nuanced solution that is not biased toward a single theoretical viewpoint. However, such multimodel frameworks require teamwork to build.

A related challenge is the assessment of models that combine separate subsystem models. Even if assessment methods may have been applied to the separate subsystems, their interaction will likely lead to new behavior. Once coupled, the full system model will need to be assessed to ensure that couplings and interactions are sufficiently realistic and capture the key features of the real-world system. Software quality checks are also helpful since errors and bugs may creep in as subsystem models are combined.

How large will the difference between prediction and reality likely be?

Uncertainty quantification is commonly associated with the task of attaching uncertainties to model-based predictions. Uncertainties arise from a variety of sources, including (1) measurement errors in the data, (2) uncertainty in model inputs (e.g., initial conditions, boundary conditions, and forcings), (3) uncertainty in model parameter settings, and (4) differences between the conceptual model and the real-world system (structural errors). A variety of approaches are used to estimate and represent uncertainty (NRC, 2012). In prediction-focused investigations, uncertainty quantification might involve a detailed analysis, using probability distributions to quantify prediction uncertainties. In more exploratory investigations, uncertainty quantification might involve a sensitivity analysis, exploring a range of possible model outcomes to understand the effect of input and parameter uncertainty. With sensitivity analyses, it is important to consider how changing multiple inputs simultaneously affects the resulting model output (Saltelli et al., 2008). Also, sensitivity analysis exploring different resolutions of modeling (if possible) can help model developers determine the level of resolution required for the investigation (Davis et al., 2008).

The impact of structural errors depends in part on the nature of the system. Nonstationary behavior is common in complex systems, particularly in social systems in which a single, random event can catalyze new system

behavior. Models that track historic data may not be able to track future observations with similar accuracy if the system exhibits such nonstationary behavior, with the system state evolving from one regime to another. In fact, calibrating the model using historic data may lead to less accurate predictions in this new regime. In such cases, models are more useful for exploring future possibilities than for making quantitative statements about prediction uncertainty. Identifying such extrapolative situations is often difficult, but it is crucial for realistic model assessment.

Estimating uncertainty due to structural error is an active area of research (NRC, 2012). In some cases, biases induced by structural errors can be corrected by applying an artificial adjustment factor, effectively adding more empiricism to the model (Bayarri et al., 2007; Kennedy and O'Hagan, 2001). Such artificial adjustments may improve forecast skill (accuracy), but they also undermine the integrity of the model. Patterns of bias depend on the state of the system, and so there is no guarantee that the bias correction will be appropriate under different conditions.

More physically motivated model adjustments for structural errors have also been proposed (Oliver et al., 2015). Perhaps the most popular approach combines results from multiple models, each of which may be consistent with the observations but yields different predictions. The spread of results from these models may better capture uncertainties about what will happen in the real-world system than any one model. Multimodel ensembles, in which different models run the same scenario using identical protocols, are commonly used in climate studies to convey uncertainty in the future climate (IPCC, 2013). This approach, which attempts to discern structural uncertainty in the model representations, is distinct from the more common ensemble methods used to capture uncertainty in inputs (initial conditions or empirical parameters).

Combined results from multiple models can also produce more accurate predictions. For example, while each of the approximately 20 climate models in the Coupled Model Intercomparison Project (CMIP) had its own biases, the prediction accuracy of the mean across all the models was better than that of any individual model (Reichler and Kim, 2008; see Figure 3.6). Related ideas appear in statistics, where shrinking noisy estimates towards the group average leads to more accurate predictions (Efron and Morris, 1977), and in machine learning, where the adaptive boosting method combines several "weak" models to produce a "strong" algorithm (Schapire, 2003).

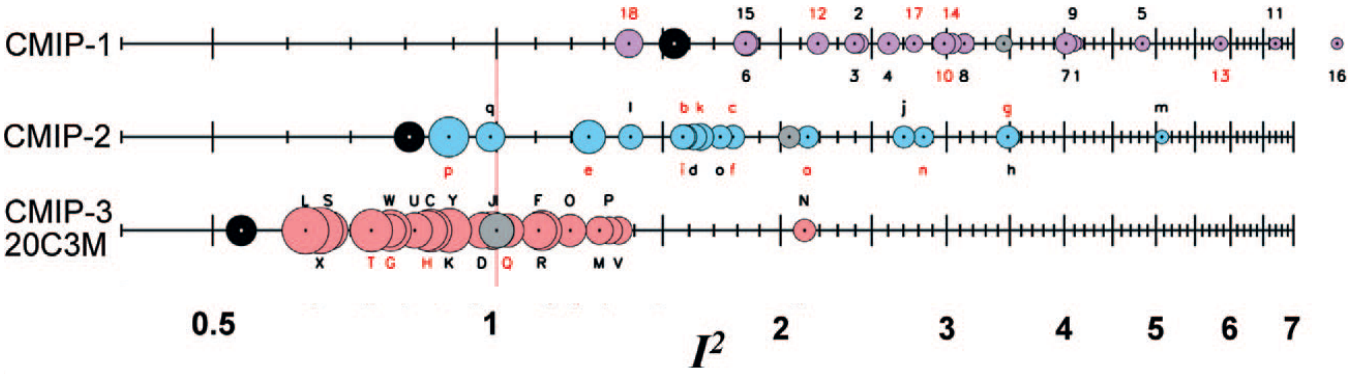


FIGURE 3.6 Increasing accuracy in simulating 20th-century climate over three generations of climate models, from the mid-1990s (CMIP-1) to 2007 (CMIP-3). Each colored circle labeled with numbers or letters represents a particular model and the black circles represent the average across all models in the multimodel ensemble. The best performing models (left) have the lowest performance index (I^2) factors. The multimodel mean generally outperforms any single model. SOURCE: Modified from Reichler and Kim (2008). © American Meteorological Society. Used with permission.

Although such model combination approaches have demonstrated some improvement in predicting past observations, there is no assurance of their benefit in more extrapolative settings, where the system state may have moved to a new regime. Indeed, all of the models may have a common bias that will never be removed by

averaging. Moreover, the effectiveness of combined predictions can be eroded by a variety of factors, such as a lack of independence, lack of diversity, and social influence (Lorenz et al., 2011).

Some important features of a model-based investigation that could affect uncertainty quantification include the following (NRC, 2012):

- The amount and relevance of the available system observations,
- The accuracy and uncertainty accompanying the system observations,
- Changes in the sensor technology used to collect data or in the way the sensor technology operates,
- Human cognitive biases in collecting and coding data,
- The complexity of the system being modeled,
- The degree of extrapolation required for the prediction relative to the available observations and the level of empiricism encoded in the model,
- The existence of model parameters that require calibration using the available system observations,
- The computational demands (run time and computing infrastructure) of the computational model,
- The accuracy of the computational model's solution relative to that of the mathematical model (numerical error), and
- The accuracy of the computational model's solution relative to that of the true system (structural error).

COMPUTATIONAL ENVIRONMENT

The computational infrastructure—including processors, software, memory, storage, connectivity, algorithms, and data structures—affects how modeling, analyses, and data can be used to support an investigation. For example, digital computing played a key role in the development of weather and climate models (see Box 1.3). The demands of models being used can also drive the requirements for computational resources. Large-scale computational models, such as the weather models mentioned above, are designed to run on compute-intensive high-performance computing systems for hours or days, running on thousands of processors, holding huge state vectors in distributed memory and storage, making use of specialized graphical processing units, and producing high volumes of model output that require specialized data storage capability (see Appendix B). At the other end of the spectrum, some system dynamics models and empirical models can be programmed quickly and run on a laptop or smaller devices.

In addition, the data and analysis requirements surrounding the model (e.g., data assimilation, inverse modeling, and sensitivity analysis) make additional demands on the computational infrastructure. For example, some analyses require a large number of computationally demanding model runs to be carried out. Whether the model runs are carried out serially or in parallel depends on the processing availability. Also, how much of the model output can be stored for later use and analysis depends on the available storage. Other data-intensive analyses involve preprocessing huge volumes of model output or system observations, which often requires special infrastructure for holding, processing, and visualizing the data.

Given these considerations, a question for NGA is the following:

What computational infrastructure is required?

A model might be computationally intensive, with code that leverages the specific connectivity and processing features of a high-performance system. It might be data intensive, with massive or streaming data, requiring specialized file and buffering systems. It might require an integrated digital environment, in which computing, data storage, and visualization systems are linked by software and high-performance networks. It may also be necessary to push data or analysis products to personal tablets or phones, making some form of cloud data or computing environment necessary. Major types of computational infrastructure are summarized below.

High-Performance Computing

Large-scale computational models of physical processes (e.g., material behavior, weather and climate, and subsurface flow) have driven the requirements and design for high-performance computing over the past half century. Traditional high-performance computing could be labeled computationally intensive, in that its design was heavily oriented toward facilitating computation (such as for partial differential equations or signal analysis) as opposed to, say, intensive throughput of data. This focus of traditional high-performance computing is evident from the LAPACK benchmarking of such systems, measuring the time required to solve large, dense systems of linear equations with little or no data input or output.

Increased resolution, inclusion of additional processes in the model, and use of growing amounts of physical observations and data products have driven advances in the size and speed of modern high-performance computing environments. Applications require a high degree of parallelism and fast communication between processors to ensure that information is quickly and appropriately transferred across spatial and temporal domains to accurately solve large, dense systems of equations and many other iterative computations. Thus, high-performance computing and compute-intensive software platforms (e.g., MPI, OpenMP) provide extremely fast facilities for coordinating tasks running on different processors. The hardware architectures (e.g., infiniband) provide fast, high-volume links connecting processors and main memories. However, the interconnection bandwidth between main memory and secondary disk storage is much lower than that among processors and main memory. This limits interactions between main memories and secondary disks to initial loading of data and storage of the final result and a few intermediate states. It also limits the volume of data that can be processed to the size of main memories, even though the secondary disk storage may hold much larger data sets.

Although many of the technologies used in high-performance computing are based on commodity components (see Appendix B), the design of the computing system is application specific. In general, a center seeking to acquire high-performance computing will design a benchmark of test cases to resemble the expected workload over the next 2 to 5 years. Standard benchmarks (e.g., High Performance LINPACK, and High Performance Conjugate Gradients)[2] are unlikely to be sufficient because there is often an appreciable gap between theoretical peak performance and actual sustained performance (referred to as percent of peak, which is commonly in the 1 percent to 10 percent range). The benchmark is used to design a system with appropriate performance levels and size for processing, input and output, networking, and storage. These different subsystems may come from different vendors and require the services of an integrator. The acquisition will also include specifications for all the needed software and libraries, including large-scale schedulers to manage the workload of many jobs on a large system.

Data-Intensive Computing

The modern deluge of data (e.g., from social media, automated transaction records, remotely sensed data, scanner data, text, and computational model output) has motivated the conception and expansion of data-intensive computing. Unlike the computationally intensive computing model commonly used for physical process models, data-intensive computing focuses on exploiting (1) massive parallelism to carry out common data analysis tasks, such as searching, organizing, aggregating, analyzing, and modeling; (2) visualizing big data with large volume, high update rate, and tremendous variety; and (3) heuristic algorithms to support data compression and to estimate global behavior from local activity.

A dominant paradigm for data-intensive computing comes from Google's MapReduce programming model and architecture (Dean and Ghemawat, 2008), which supports high-bandwidth communication between main memories and secondary disk storage. Apache Hadoop, an open-source implementation based on MapReduce, is commonly used by companies leveraging big data. The Apache Hadoop project has evolved to include additional

[2]See http://hpcg-benchmark.org.

subprojects that bolster processing and analysis capabilities; these include Hive, Pig, and Sqoop (see Figure 3.7). These paradigms work especially well for problems that can be easily parallelized, such as text mining. For complex networks, however, these paradigms are less effective because their use often requires heuristic compression techniques, which eliminate key structural elements. Moreover, for streaming data where iterative network analytics are needed (such as analyzing the Twitter firehose), the MapReduce tasks are too coarse grained. In such cases, a SPARK approach or techniques that use the graphics cards or specialized chips designed for network metrics may be more effective. Finally, when data volume is so massive that it requires a cluster of machines and interactive queries of the data, the overhead of running MapReduce will be noticeable.

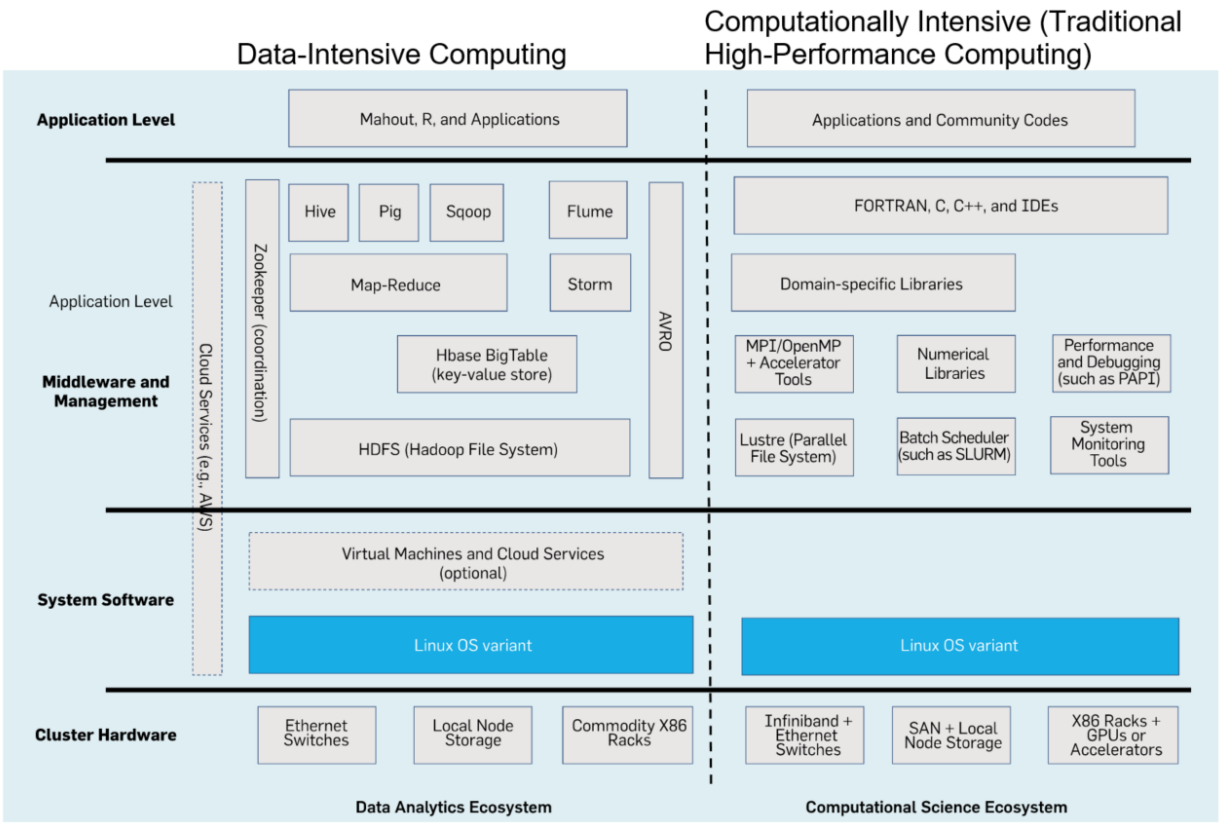


FIGURE 3.7 Two main supercomputing paradigms: data-intensive computing (left), and computationally intensive high-performance computing (right). Data-intensive computing predominantly uses the MapReduce paradigm, parallelizing the work into separate tasks that require little or no communication or coordination between them. Traditional high-performance computing relies on MPI, numerical libraries, and centralized storage. SOURCE: Reed and Dongarra (2015). © 2015 Association for Computing Machinery Inc. Reprinted with permission.
NOTE: AVRO is Apache Software Foundation's remote procedure cal and data serialization framework; AWS = Amazon Web Service; FORTRAN = Formula Translation; GPU = Graphics Processing Unit; IDEs = Individual Development Environments; MPI = Multi-Point Interface; Open MP = Open Multi-Processing; OS = Operating System; PAPI = Performance Application Programming Interface; SAM = Storage Archive Manager; SLURM = Simple Linux Utility for Resource Management.

The MapReduce paradigm carries out operations on data residing on a distributed file system by distributing a task over the data set, carrying out the tasks locally on the separate data pieces, collecting the distributed intermediate results, then producing the final result. This approach is efficient for embarrassingly parallel tasks,

where there is no need to share information between the operations being carried out on the separate pieces of the data until the operations are complete. Such operations have proven useful in text mining applications where summaries of massive text or data sources are created to serve as raw material for analyses. This paradigm is not efficient for iterative applications because separate, distributed operations will generally need to share information regarding nearness and dependence. New data-intensive software platforms (e.g., Apache Spark) are emerging to improve performance of iterative tasks, which are common in spatial and compute-intensive applications. These new platforms are better for network models and for social media analytics.

Spatial Computing

Spatial computing refers to computing in spatial, temporal, and spatiotemporal spaces across both geographic and nongeographic domains (e.g., indoor spaces). Models and data with spatial or space-time-dependent structures do not fit easily into either data-intensive or traditional high-performance computing-based analysis frameworks for a number of reasons, such as the following:

- It is difficult to divide many spatial analysis tasks into equal subtasks that impose comparable costs to different processors because of spatial data diversity and spatial variability in data density. Data density (e.g., number of houses per unit area) varies substantially from rural to urban areas, and so standard data division techniques (e.g., random, round robin, hashing, and geographic space partitioning) are not ideal.
- The computational load associated with spatial data types (e.g., line strings, polygons, and polygon collections) varies greatly with the shape, spatial context, and query restrictions of the elements, thereby reducing the effectiveness of standard data division techniques. In addition, it is often more expensive to move a data element (e.g., a polygon with thousands of edges) than it is to process it locally using filter-and-refine approaches that leverage spatial indices and minimum orthogonal bounding rectangles. This makes common dynamic load balancing and data-partitioning schemes inefficient (Shekhar et al., 1996, 1998).
- Many spatial computations (e.g., shortest path computation and parameter estimation for spatial autoregression) are difficult to decompose into independent tasks suitable for data-intensive computing architectures. Achieving optimal spatial decomposition for load balancing has been proven to be difficult computationally (Wang and Armstrong, 2009).

The need for efficient manipulation and scalable analysis of spatial big data on distributed archives has spawned a vigorous research effort focusing on spatial modeling within compute- and data-intensive cyberinfrastructures. This effort has led to an increasing array of new technologies (e.g., geographic information system [GIS] Tools for Hadoop, SpatialHadoop, cyberGIS, and GABBS) exploiting spatial characteristics and parallelism of both data and computation. High-performance networks, such as those at multiple-hundred gigabytes across distributed archives, are needed for data-intensive modeling.

The development of cyberinfrastructure and cyberGIS capabilities could enable both compute- and data-intensive geospatial modeling (Wang et al., 2014). Tools for geospatial problem solving and decision making are typically designed for individual user groups and there is limited coupling between various data and models. A digital discovery environment would provide: (1) a well-defined set of computational and data objects and services for model-based investigations and simple ways of combining them in complex coupled models, and (2) user-friendly and interoperable ways to add new data or computation objects and services, and to allocate cyberinfrastructure resources and services. The cyberGIS community has made solid progress on developing such an environment to meet the needs of various geospatial communities (Wang et al., in press).

Cloud Computing

The National Institute of Standards and Technology defines cloud computing as an approach "for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" (Mell and Grance, 2011). Data access, delivery, and computing power from shared computational resources are supplied as required by the users. Common uses of cloud computing include email, Dropbox, Google drive, and online search engines.

The computational infrastructure behind the cloud is primarily data-intensive computing. For such applications (e.g., mapping), distributed content delivery networks allow data to be pushed out to large numbers of users as necessary. Some hosted high-performance resources are also available from the cloud (e.g., Amazon Cloud and the U.S. Department of Defense High Performance Computing Modernization Program). Such systems are able to supply the necessary compute cycles for high-demand users, and they may also facilitate sharing of large spatial data sets with users that have limited computational, storage, and communication resources. For example, NGA and the Polar Geospatial Center are increasing access to high-resolution polar imagery from Digital Globe satellites using cloud computing platforms at the University of Minnesota and National Center for Supercomputing Applications.[3]

With cloud computing, users pay only when they need it, and there is no need to purchase and maintain computing infrastructure. However, security and reliability have to be handled differently than in-house high-performance computing. A somewhat recent assessment of cloud computing for science appears in the report of Yelick et al. (2011).

Whether the supercomputing resources are hosted on the cloud or in data centers, the architecture needs to accommodate the computing requirements of the application. For example, many data-mining applications can leverage the MapReduce model, using embarrassingly parallel programs that can be hosted on huge server farms such as Amazon or Microsoft Azure. In contrast, many physical process models require highly parallel, computationally intensive high-performance architectures. Consequently, as new modeling and analysis approaches are considered, it will be helpful to ask the following:

Should the modeling and analysis leverage special features of the cyberinfrastructure?

Generally, models and analyses that explicitly leverage features of the available cyberinfrastructure have greater capacity, capability, and speed. For example, climate models that use cutting-edge high-performance computing can run at higher resolution or include more detail in the modeling, perhaps yielding results that are more like the real-world system of interest. Similarly, analysis approaches that use data-intensive computing can handle a much larger data volume and rate for constructing empirical models or carrying out inference, potentially producing faster and more comprehensive results.

For applications that require real-time computation in the field (e.g., models used to support battlefield decisions), limitations on both computational power and data access have to be overcome. The load on the communication channel to centralized cloud computers can be reduced by enhancing the computing power on board sensor platforms such as unmanned aerial vehicles and aircraft. Specialized high-performance computing platforms (e.g., FPGA and GPU clusters) and custom software (e.g., signal processing algorithms designed for FPGA) provide a means to increase communication without exceeding weight or power restrictions on aerial platforms.

Adapting data, models, and analyses to specialized access modes and cyberinfrastructure has costs. In addition to the cost of hosting the infrastructure or "renting" it via cloud services, model development and use is generally more difficult and takes longer, and the infrastructure has its own maintenance requirements. Also, models must evolve with the infrastructure. Consequently, the utility of adapting the model and methods to new access modes

[3]See https://www.whitehouse.gov/blog/2015/09/02/using-new-elevation-data-explore-arctic.

and supercomputing architectures will depend on features and commonalities of the investigations NGA will carry out.

TRADEOFFS

A key part of any model-based analysis is managing tradeoffs between the various features of the available models and analysis methods, balancing the available resources to address the key questions of the investigation. Table 3.1 lists examples of resources and features of the real-world system that must be balanced in any model-based investigation.

TABLE 3.1 Example Factors Affecting the Balance Between Resources, Key Questions, and Features of the Real-World System

Available Resources	Key Questions and Features of the Real-World System
• Expertise	• Complexity of the real-world system
• Labor/manpower	• Do the key questions center on predictions, or do they center on exploration of plausible outcomes?
• Time	
• Existing models	• Are the key questions "extrapolative" or captured in past data and experience?
• Data	• Amenability of the real-world system to modeling to address key questions
• Computation	• Accuracy of the result required

A key property of a model's impact on the investigation is its fidelity to the real-world system being modeled. Higher-fidelity models capture a larger number of important subsystem behaviors and linkages present in the real-world system, with greater mathematical and computational accuracy, and they are also typically higher resolution. In contrast, low-fidelity models are crude, sometimes empirically constructed, yielding simple representations of only a few system processes (e.g., the left frame of Figure 3.2). Data and computational requirements for high- and low-fidelity models can be vastly different. Higher-fidelity models generally require more data and at a higher level of detail, more sophisticated analysis methods to bring the model in line with the real-world system, more demanding computations to run the models and carry out analyses, and more time and expertise for assessment.

Lower-fidelity models can produce accurate predictions in some situations, such as when the real-world system is dominated by a small number of processes. When system behavior becomes more complex, with contributions from multiple interacting processes, high-fidelity models are required to produce realistic simulations. This additional realism is particularly important for characterizing uncertainty in extrapolative settings, where the model must explore plausible, uncertain future outcomes.

From the perspective of empirical models (i.e., statistics and machine learning), fitting additional features in the data affects the error in the resulting model predictions. The error can be decomposed into bias and variance. The bias is the mismatch between the model and reality, and the variance is a measure of the size of random fluctuations in the data. More complex models can be overfit to the training data, following the random fluctuations in the data. Managing the complexity of an empirical model depends on the bias–variance tradeoff, a well-covered topic in the literature (e.g., Hastie et al., 2009). Consequently, a key question is the following:

How do modeling tradeoffs affect fidelity to the real-world system?

Answering this question requires balancing the quality of the answer with the time and other resources available to obtain it. If time for the investigation is short, the available resources can be marshaled to produce a quick,

approximate answer, rather than a more complete answer months or years later. If ample time is available, a greater investment in expertise, high-fidelity modeling, data collection, and analysis methods can be made to address and explore key questions.

In addition, the features and requirements of the investigation can help determine the appropriate level of fidelity. For example, if specific, highly accurate predictions are required (as for the wave prediction model in the pirate example of Chapter 2), then a commensurate level of model fidelity and data will be required. Similarly, more extrapolative or exploratory investigations will likely require higher-fidelity modeling, as well as sufficient computing resources to carry out extensive sensitivity studies.

SUMMARY AND CONCLUSIONS

In a model-based investigation, the key questions should drive the focus of the analysis (e.g., prediction or understanding) and the type of model needed (e.g., process or empirical). For NGA, which often must produce geospatial intelligence quickly, timeliness of results is also a consideration for model selection. A model based on a simple, quick approach may be more useful than one that is more comprehensive but slow. In such cases, it may be possible to utilize existing models or model output, which would both speed the investigation and reduce its cost. However, NGA would have no control over, and perhaps insufficient information about, what went into the model. Building new models from scratch or from a combination of subsystem models would allow NGA to target the behaviors and processes of interest, but it would also require significant time, effort, and expertise. Because models of complex systems are highly multidisciplinary, experts will commonly have to be drawn from modeling groups across the country.

The need to work in multidisciplinary teams of experts is particularly important when combining models, because the robustness of the results depends on getting both the subsystem models and the model linkages right. Multiple subsystem models are often needed to represent the relevant features of a real-world system. However, simply combining models that accurately capture subsystem behaviors will not yield a model that accurately captures the behavior of the larger system, unless the connections among subsystem models are appropriately represented. Combining models is challenging but would give NGA some flexibility in model development. In particular, a collection of subsystem models that could be used in a variety of larger models would be adaptable and less expensive to develop and maintain than a few megamodels.

All models are abstractions of real-world systems, and they must be assessed to determine their adequacy in reflecting real-world behaviors, in the context of supporting specific decision making. How the assessment is carried out depends in part on the purpose of the model. For example, a model used to make high-consequence, specific predictions (e.g., the load of a beam in a structure) requires rigorous validation and uncertainty quantification procedures so users can determine how much trust to place in the predictions. In contrast, a model built to explain phenomena may undergo more qualitative procedures for the validation assessment. However, even a largely qualitative model assessment would be useful for communicating results, strengths, and limitations of a model-based analysis. Such communication is particularly important when the users are not involved in the modeling, as is the case for decision makers who use the geospatial intelligence produced by NGA.

Likewise, it is important to understand the sources of uncertainty in the investigation and to communicate the overall level of uncertainty to users. When relevant historical data are available to repeatedly test model-based predictions against the real-world system outcome, there are a variety of methods available to estimate and represent uncertainty. In contrast, it is difficult to quantify uncertainty in models of systems with processes that are evolving or changing to a new state. Methods for dealing with these uncertainties and for communicating uncertainty in model results are active areas of research. If NGA uses an existing model, it may be possible to work with the model developers to understand how uncertainty is represented or even to have uncertainty expressed in a way that is useful to NGA. In addition, if a collection of different models and methods produces similar results, NGA may have higher confidence in those results.

The nature of the data (e.g., how it was collected, sources of errors, accuracy, and volume) available for a model-based investigation influences model selection and drives how it is handled in the analysis. For example, it may be preferable to use simple, data-driven models if high-quality data that are directly relevant to the questions of interest are available. Specialized cyberinfrastructure and big data techniques are needed to handle large volumes of geospatial data collected at different scales from different sources and for different purposes.

The computational infrastructure needed depends on the type of model being run. In general, large-scale, physical process models use traditional high-performance computing for models that require large numbers of processors, fast communication between processors, and large volumes of data in memory and storage (e.g., climate models); they use data-intensive computing for models that analyze massive amounts or streaming data by dividing the work into separate, parallel tasks (e.g., preprocessing of tweets in the Benghazi consulate example in Chapter 2). Cloud computing is a potential pathway for accessing the computational infrastructure—data-intensive or traditional high-performance computing—required for the modeling and analysis effort, as well as providing on-demand data or analysis products (e.g., online search engines). Specialized software platforms are usually required for large-scale models with spatial dependence, because the analysis tasks (e.g., hotspot analysis) cannot be easily parallelized. In general, models that explicitly leverage available computational infrastructure have greater capacity, capability, and speed, although they are more costly to develop and maintain.

4

Models and Methods Relevant to NGA

Chapter 3 described ways to think about the models, data, analysis, and computation necessary for a model-based investigation. This chapter covers the committee charge (see Box 1.1), including a description of types of models and methods (Task 1), their relevance to the National Geospatial-Intelligence Agency (NGA) (Task 2), the state of the art (Task 3), and actions and research needed to make them more useful for geospatial analysis (Tasks 4 and 5). The chapter begins with a discussion of how the committee addressed each task. The remainder of the chapter is divided into sections on different classes of models and methods, each of which covers all of the tasks.

Given the breadth of national security and humanitarian challenges under NGA's purview, it could be argued that dozens of models and analysis methods, each with important variants, are potentially relevant to NGA. It is not possible or useful to discuss every one of them to address Tasks 1 and 2. Instead, the committee focused on broad categories of models and methods that connect directly to NGA's mission and that would help address the two example intelligence scenarios provided by NGA. With regard to the first, NGA's mission is to produce geospatial intelligence by assessing and visually depicting physical features and geographically referenced activities on Earth. To extend this task to the modeling realm, NGA will need models of human behavior (activities), set within an environmental context (physical features), to develop scenarios and make predictions, as well as techniques for integrating, analyzing, and verifying geographically referenced data in a model-based investigation. These needs place a premium on the following types of models and methods:

- Models of physical processes that affect human activities (e.g., weather and water flow);
- Social system models of human behavior in a geospatial context;
- Models of combined physical and social systems;
- Inverse methods to infer uncertain model parameters from measurements of the real-world system;
- Spatial statistics, data mining, and machine learning to discover trends, patterns, and associations in disparate data; and
- Spatial network analysis to examine how patterns of relations affect behavior at the individual to state level.

The sections below give examples of how these models and methods would contribute to answering the questions posed in the intelligence scenarios provided by NGA:

1. Megacities (greater than 10 million people): *How will worldwide urbanization trends affect regional political, economic, and security environments?*
2. Chinese Water Transfer Project: *How do agriculture and energy production and consumption change over time? How and where will populations, including rural communities, shift?*

Task 3—a description of the current state of the art in models and methods, including features and scales captured by the model, accuracy, reliability, predictability, uncertainty characterization, and computational requirements—was difficult to address for two reasons. First, each of these factors is a product not only of a model, but also of the particular context in which it is being used. As discussed in Chapter 3, the key questions driving the investigation will influence what models are used, the processes and features represented in the model, the type and accuracy of data needed, how or whether the model's fidelity to the real-world system will be assessed, and the computational requirements. Thus, a description of the state of the art for an individual model is unlikely to apply to all variants and applications of that model, and even less likely to apply to a category of models. Second, many of the factors specified in Task 3 (e.g., scales and predictability) do not apply to analysis methods, and other useful factors (e.g., availability of software, training support, and data issues) are not specified. Consequently, the committee developed a common set of state-of-the-art factors for all of the models and methods discussed in this chapter, and often provided ranges or examples to describe them.

For Task 4, the committee considered actions NGA could take to use or adapt existing models, given the educational profile of NGA analysts (NRC, 2013), NGA's experience with modeling, the difficulty of developing or reusing relevant models and methods, the availability of software and code, and the level of training support available. The objective was to identify a short list of ideas for making existing models more useful to NGA, not to produce a comprehensive action plan. Many of the actions could be undertaken in collaboration with partners in universities, federal agencies, and private companies. NGA already has relationships with a number of universities that have strong programs in geospatial science (see Box 4.1), some of which also have experience in models and methods discussed in this chapter.

Research funded by NGA offers another clue about NGA analysts' knowledge and experience and also provides guidance on future research and development needed to use the models and methods for geospatial intelligence (Task 5). For example, spatial and temporal analyses have been major research themes for NGA for at least the past decade, and algorithms for data-intensive computing, particularly for image analysis, have been a research theme for the past 5 years.[1] Consequently, NGA likely has reasonable capacity and connections with outside experts in these areas. In contrast, space-time modeling and predictive models have only recently become research areas for NGA, and so capacity and connections will likely have to be developed.

PHYSICAL PROCESS MODELS

The physical system serves as the environment in which the social system evolves. Physical process models are executable descriptions of our understanding of atmospheric, oceanic, hydrologic, geologic, and other physical systems. Many of these processes lend themselves to geospatial analysis. Physical process models developed or used by NGA include those used to generate high-resolution representations of Earth's magnetic and gravitational potential (Pavlis et al., 2012; see Figure 1.2). While grounded in the simulation of specific natural phenomena, physical process models often also supply information on the impact of environmental dynamics on human infrastructure, activities, and demographics. For example, the megacities intelligence question (see Box 1.2) would

[1] See NGA Academic Research Program Symposium programs for 2005–2015.

BOX 4.1 NGA Partnerships with Universities

NGA has established relationships with dozens of colleges and universities, including historically black colleges and universities, for recruiting and continuing education purposes (NRC, 2013). In addition, NGA has selected a large number of universities as Centers of Academic Excellence in Geospatial Science as a means of cultivating relationships and partnerships.[a] These include the following:

Alabama A&M University
Arizona State University
Delta State University
Fayetteville State University
George Mason University
Mississippi State University
Northeastern University
Ohio State University
Pennsylvania State University
Roane State Community College
University of Alabama
University of Maine
University of South Florida
University of Texas, Dallas
University of Utah
U.S. Air Force Academy
U.S. Military Academy

[a]See NGA Academic Research Program Symposium programs for 2005–2015.

likely require models of environmental changes that could stress urban populations, such as sea-level rise and increases in summer temperatures. The Chinese water transfer questions would likely require a large-scale model of the hydrologic system in China to predict surface flow, subsurface flow, and abundance of water under different water diversion scenarios.

Many physical process models involve simulating fluids that are governed by Navier-Stokes and continuity equations, which represent conservation of momentum and mass, and they are solved numerically through finite or spectral discretization approaches. Accurate and representative observations of the natural system are critical both for creating the physical process models themselves and for setting the correct initial and boundary conditions that constrain the physical processes in a model-based investigation.

Physical process models can be large, highly nonlinear, and may couple together multiple processes over a wide range of space and time scales (e.g., Figure 4.1). Large, complex physical process models are often expensive to develop and run. In some cases, it may be sufficient to run reduced-order models, which use theoretical approaches to develop a simplified version of the full process model (Berkooz et al., 1993; Mignolet and Soize, 2008; Moore, 1981). Reduced-order models are intended to provide adequate approximations to high-fidelity models at significantly lower cost and time-to-solution.

A reduced-order model need not be faithful to the full spatiotemporal dynamics of the high-fidelity model; it need only capture the essential structure of the simulated structure from the input parameters to the outputs of interest. The most popular approach to model reduction is to reduce the state dimension and the state equations using projection-based methods (Benner et al., 2015; Chinesta et al., 2016). Such methods are most successful for linear or weakly nonlinear models in low parameter dimensions. However, constructing efficient and capable reduced-order models that can handle complex nonlinear dynamical models and that are faithful over high-dimensional parameter space remains challenging.

Reduced-order models or emulators, which replace the computational process model with a response surface model, are also used to speed inverse or sensitivity analyses that require a lot of model runs. The emulator is trained from an ensemble of physical process model runs and creates a response surface mapping model inputs to model

output (Marzouk and Najm, 2009; O'Hagan, 2006). Emulators can be used to predict model outcomes at untried input settings, allowing more thorough inverse or sensitivity analyses to be carried out.

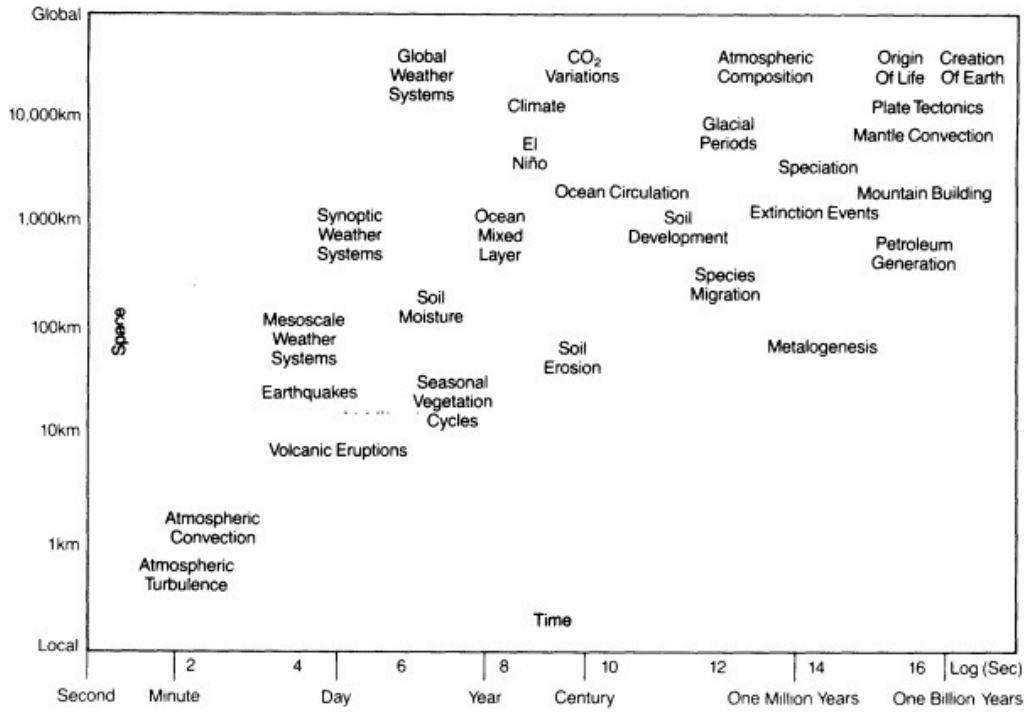


FIGURE 4.1 Characteristic spatial and temporal scales of Earth system processes. SOURCE: NAC (1986).

State of the Art

Some physical processes, such as turbulence and heat and fluid flow, are relatively well understood and modeled due to a wealth of observation and theory. Other physical processes, such as subsurface dynamics, clouds, and ocean biogeochemistry, remain a challenge to model because of a lack of observations. This is a particular problem for highly detailed simulations of flows or other processes that vary at fine spatial or temporal scales (e.g., groundwater flows and precipitation). In addition, there are uncertainties regarding how any natural system responds to forcings and feedbacks.

Climate models are mature for the questions and scales they were designed to address and are a good example of the state of the art in physical process models. The models are complex—involving multiple processes (see Figure 3.4) and multiple types and scales of observations—are computationally demanding, and typically require a dedicated research center or a large community of researchers to develop, validate, and run. A global research community has emerged to simulate consistent past, present, and future climate-based scenarios of climate drivers, such as greenhouse gas concentrations, volcanic and manmade aerosols, and solar strength. Although the response of individual models to climate forcings varies, uncertainty is typically minimized by using multimodel ensembles that have run the same scenario (e.g., Figure 3.5). One of the greatest uncertainties in climate models concerns which emissions scenario will match future societal choices about energy, transportation, agriculture, and other factors (IPCC, 2013).

Substantial efforts have been made to downscale large-scale climate simulation results to the regional or local

scales that are more relevant to decision making (Kotamarthi et al., 2016). The goal of downscaling is to achieve the realistic high-frequency spatial and temporal variance of the real world that the coarser information lacks. There are three primary downscaling approaches:

1. Simple, which adds trends in the coarse-scale data to existing higher-resolution observations (Giorgi and Mearns, 1991);
2. Statistical, which relates large-scale features of the coarse data to local phenomena using regression methods, typology classification schemes, or variance generators that add realistic high-frequencies to the coarse data (Wilby et al., 2004); and
3. Dynamical, which uses the coarse information as input to high-resolution computer models to dynamically simulate phenomena at much finer temporal and spatial scales.

Each downscaling approach has its own advantages and disadvantages. The less complex methods are simple, fast, and inexpensive to calculate, but they may produce inaccurate results, particularly for future scenarios that may differ from historic observations (Gutmann et al., 2014). Dynamical downscaling is more complex and expensive, but it has the potential to generate high-resolution information over a wider range of extrapolative scenarios. An example of a downscaling application is illustrated in Figure 4.2.

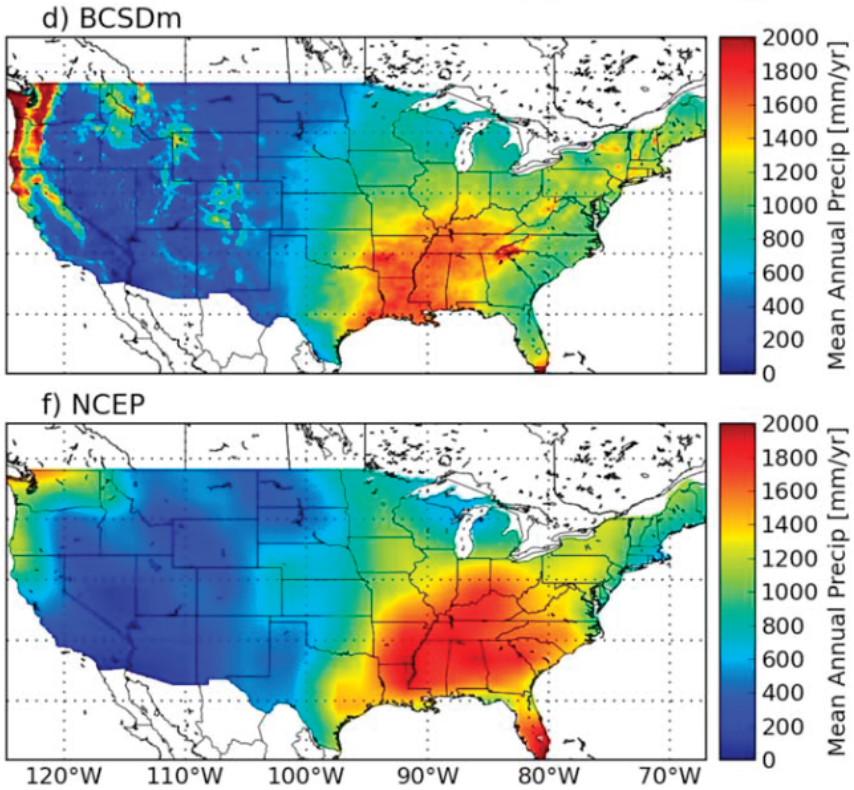


FIGURE 4.2 Mean annual precipitation downscaled (upper) from NCEP-NCAR Reanalysis (lower) helps capture the regional influences of high variability terrain, such as the Rocky Mountains. NOTE: BCSDm = bias corrected spatial disaggregation, applied at a monthly timestep, the statistical method used for downscaling. NCAR = National Center for Atmospheric Research; NCEP = National Centers for Environmental Prediction. SOURCE: Gutmann et al. (2014).

The state of the art in physical process models, particularly weather and climate models, is summarized in Box 4.2.

BOX 4.2 State of the Art in Physical Process Models

- **Space and time scales:** The scales simulated by the family of physical process models cover subatomic to galactic spatial scales and near-instantaneous to geologic time scales. Specific physical process models and methods are generally designed to simulate specific processes and features within a limited space and time scale (see Figure 4.1) and are rarely valid outside of those scales.
- **Fidelity:** Physical process models are designed to balance the degree of fidelity to the amount of accessible computing power, the time available to carry out the simulation, and the complexity of the system being modeled. Models range from simple, low-fidelity models that can be run in seconds on a low-end tablet computer to very complex high-fidelity simulations that take months to run on the fastest supercomputers.
- **Accuracy and precision:** In physical process models, accuracy refers how closely the simulation matches the average behavior of the observed system, whereas precision is a measure of the variance. A prediction that tomorrow's temperature will be less than 200ºF is surely accurate but imprecise, whereas a prediction that it will be 23.456789ºF is precise but likely to be inaccurate. The accuracy of many types of physical process models, such as numerical weather models, has been improving over time because of constant advancements in scientific knowledge and technology (Bauer et al., 2015). Both accuracy and precision are taken into account in measures of model prediction skill.
- **Predictions and scenarios:** The predictability limit for an individual forecast of the highly nonlinear weather system is about 2 weeks. The predictability goal for climate models is currently a season to a decade (a) through better knowledge and observations of the ocean thermodynamics, which is the principal driver of the system at those scales, and (b) by forecasting outlooks of lower-precision features, such as wet or dry trends over broad areas, rather than precise temperatures or rainfall amounts at specific locations (Slingo and Palmer, 2011).
- **Uncertainty analysis support:** In physical process models, uncertainty characterization is a method for conveying the uncertainties that are inherent when simulating continuous environments with discrete grid points and time steps, using imperfect observations and models. Because physical process models are increasingly being used in decision-making contexts, more emphasis is being placed on quantifying model uncertainty in a manner that makes the data more usable and actionable.
- **Validation and assessment support:** Process models are validated though detailed comparisons of the accuracy and precision of the simulations relative to the observations of the physical system being simulated.
- **Computational requirements:** State-of-the-art physical process models have matched their computational requirements to the rapid increase in computational capability over the past two decades. Petascale (1015 floating-point operations per second [FLOPS]) computers became firmly established in 2014, and exascale (1018 FLOPS) architectures are currently being designed.
- **Data requirements:** Physical model data output ranges from insignificant to overwhelmingly large, even in installations with dedicated automated high-performance mass storage systems. Substantial model output is publicly available, and much of it uses standardized data and metadata formats to improve interoperability.
- **Difficulty to develop:** Low-fidelity models of simple physical systems can be trivial to develop, whereas high-fidelity models of complex systems require years of effort by large teams of researchers.
- **Reuse:** Physical process models are usually designed to be reused extensively.
- **Software/code availability:** Although software and codes for physical process models developed in academia are often open source, most physical process models developed for commercial, classified, or emerging research applications require establishing contractual relationships to acquire or use.
- **Training support:** Highly variable. While many physical process models include sufficient documentation on their use and application, others do not.

How to Make Useful for NGA

Rather than attempting to build in-house expertise in all relevant physical processes, NGA could leverage existing expertise in other organizations, either by becoming a user of model results or by becoming a partner in teams experienced in designing, carrying out, and analyzing physical process simulations. Vast amounts of process model output data are readily available, although much of it would require additional context to be useful for NGA applications, and much would have to be downscaled to the regional and local scales most relevant to NGA questions. In addition, some features of process model simulations may be reliable and useful in new applications. For example, framing the analysis in terms of risk (a function of vulnerability, exposure, and hazard) can be useful for examining the impact of physical processes on human systems. Finally, some process modeling teams develop benchmark scenarios (e.g., future emissions trajectories for climate models), and they may be willing to work with NGA to develop, run, and interpret scenarios tailored to NGA's specific interests. In all of these situations, NGA will need to invest time identifying domain experts and collaborating with them to design new simulations or scenarios, to select existing model output appropriate to the NGA question under consideration, to minimize uncertainties associated with downscaling, or to understand the strengths and weaknesses of the models in NGA scenarios.

The results of physical process models are increasingly being adapted for use in geospatial tools such as geographic information systems (GISs), which could facilitate their use in geospatial intelligence. The capabilities and sophistication of geospatial technologies as well as the large size of the GIS community have prompted many physical process modeling groups to ensure that their model results can be integrated into the rapidly proliferating suite of open and commercial GIS tools. Physical process model data can be made GIS compatible by using controlled vocabularies, standardized conventions for time and geolocation, and metadata. Once the georeferenced physical process model data are in GIS-ready formats, they can easily be mapped into human systems such as populations, cities, infrastructure, land forms, or social entities. This capability enables interactive data exploration, analysis, visualization, and distribution, all of which would improve delivery of usable model information to a broad range of users and uses (Wilhelmi et al., 2016). An example of a GIS analysis of climate model results is shown in Figure 4.3.

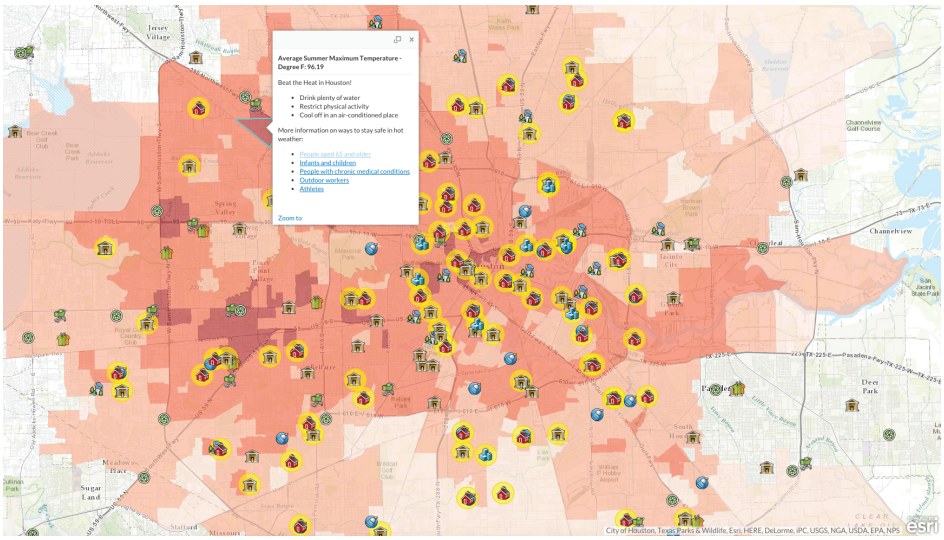


FIGURE 4.3 "Beat the Heat in Houston," a Web-based tool that integrates temperature data with information about cooling centers and Centers for Disease Control and Prevention–based recommendations for at-risk populations. SOURCE: Courtesy of Jennifer Boehnert, National Center for Atmospheric Research.

NGA-Funded Research and Development Areas

Geospatial intelligence is based on analysis of the environmental context, relevant natural and human factors, and potential threats and hazards.[2] Because so many physical processes are relevant, and because the linkages between physical process models and downstream impacts are often highly nonlinear, NGA will be challenged to determine which physical process models are sufficiently important to develop and maintain for geospatial intelligence applications. Moreover, the important physical process models will depend on the intelligence application, and their importance can only be unambiguously assessed after a complete end-to-end system has been constructed. That said, some areas do lend themselves to NGA research and development, such as the following:

- Precision real-time weather predictions to support applications such as (a) ground and aerial insertion, mission supply, and disaster relief missions in complex natural environments, or (b) source location, dispersion rate, and damage estimations from the release of hazardous materials in densely populated urban environments;
- Improved simulations to anticipate regional extreme events and environmental threats—such as droughts, relentless heat (sustained high heat and humidity) events, disease vector precursors, or rapid sea-level rise—that could lead to large-scale social unrest, disruption, or migration;
- System emulators and reduced-order models of large complex physical systems—such as climate, water, or chemical processes—to allow rapid exploration of scenarios of interest to NGA;
- Methods to rapidly and inexpensively predict the importance of physical processes on intelligence applications;
- Design of robust frameworks, couplers, and application program interfaces to include physical process models in intelligence applications; and
- Refinement of physical models to facilitate their combination with social system models to gain additional understanding and potentially predictive capability of relevant coupled physical–social system stresses and behaviors.

To help determine what physical process models are needed, NGA could survey its analysts and customers on gaps in data, information, knowledge, and capability. NGA could then build the in-house disciplinary knowledge and expertise to develop models or use available physical process data needed to fill the gaps. Partnering with other agencies that have relevant experience or mission responsibilities would likely speed the development of NGA's in-house capabilities.

SOCIAL SYSTEM MODELS

Social system models are used to understand human behavior and to make decisions in both idealized and real-world settings. Many of these models provide a means for understanding the social consequences of the diverse feedbacks inherent in a complex social system. Using these models to reason about how different courses of action will affect behavior or how different scenarios are likely to unfold will often be more important to NGA than forecasting. For example, the megacities intelligence question (see Box 1.2) would likely require models to determine what changes in social systems (e.g., financial, cultural, ethnic, religious, and health) may trigger political, economic, or security problems across different geoterrains. The Chinese water transfer questions would likely require social system model scenarios of how affected populations are likely to respond to dam construction and involuntary migration.

A variety of social system models can be made useful for NGA, but those that most naturally lend themselves to

[2]See www.nga.mil/ProductsServices/GEOINTAnalysis.

reasoning about geospatially embedded social systems are (1) system-level models and technical graphical methods (e.g., Bayesian influence models and Petri-nets) that support modeling of system-level flows and influences and (2) agent-based approaches that support modeling of emergent population behavior. These types of models support what-if reasoning, and they can be used to help NGA think through possible social responses to events embedded in particular geographic conditions, frame arguments, describe the interplay of complex processes, explain a behavior outcome to a physical process, or discuss alternative courses of action. These classes of models are described below.

System-level models. In system-level models, such as system dynamic models (Forrester, 1961; Sterman, 2000), causal loop or influence networks are used to convey how one part of the system influences another (the "wiring diagram"). Geospatial factors are incorporated in system-level models in three ways. First, direct causal and influence relations can be represented. For example, a reduction in rainfall reduces crop yields and increases food prices, and the resulting food shortage results in more illnesses and emigration. Second, geographical differences in sociopolitical influence can be captured in models of different locations (e.g., drought may be modeled as leading to emigration in one city and to more wells being built in another city). Third, changes in geographical resources can be represented in terms of stocks or the likelihood of events (e.g., level of deforestation or likelihood of building a well). The wiring diagram is often the most valuable aspect of a system-level model, because it allows the developer and the end user to gain a sense of system behavior, evoke discussion, and create common understandings. The results of the simulation itself are not always credible because of (1) insufficient theory, data, or time to instantiate the model and test the connectors between components, or (2) uncertainties in the functional form of a relation of one stock to another in the model.

Agent-based models. In agent-based models (Bonabeau, 2002; Gilbert, 2008), heterogeneous agents are placed in a context and group-level outcomes emerge from the interactions among these agents as they operate under various rules of behavior, cognitive processes, and modes of learning. Agent-based models fall into three major classes: cognitive, population, and network based. Cognitive models use only a handful of actors and focus on the detailed behavior of the actors engaged in specific tasks. The modeling frameworks often take into account differences in how data are sensed (e.g., by touch or sight), memory considerations, or time for precognitive and cognitive functions. Population-based models use large numbers of actors with simple information-processing behaviors, and they concentrate on emergent group-level phenomena. The modeling frameworks often take into account differences in the resources, goals, capabilities, or information available to the agents. Agent-based dynamic-network models (Carley et al., 2009) use a moderate number of actors with moderately complex cognitive functioning and a position in one or more social networks, so that what they know is related to whom they interact with, and the distribution of action and information changes as the social networks evolve, and vice versa. Because agent-based dynamic-network models incorporate social cognition (knowledge about groups, generalizations based on group membership, and presence of a generalized other), their predictions about social behavior are more realistic than those of other agent-based models. For all agent-based models, there is a tradeoff between the number of agents and the fidelity at which the actor's cognition and social position are modeled.

Geospatial factors in agent-based models can be accounted for in three ways. First, agents can have geographically specific behavioral rules (e.g., agents representing Europeans might have a different caloric intake and response to violence than agents representing people in the heart of Africa). Second, types of agents may be differentiated by geographical properties (e.g., agents representing global companies might have a global reach, whereas agents representing a worker might have only a city-level reach). Third, the agents can be designed to move through a virtual landscape that represents the physical geography and to behave differently at different locations in that landscape (e.g., they can only get water at a well). The landscape is often represented as a grid or toroid around which agents move, and maps may be overlaid. Models that use actual maps and direct agents to move accordingly, such as along roads, typically assume the social network is fixed or dictated by location. In

contrast, models that have better representations of the communication space (including the digital landscape) often have primitive representations of the physical landscape.

It is often more complicated to set up an agent-based model than a system-level model. Moreover, whereas system-level models can provide insight even at the wiring diagram level, agent-based models generally need to be instantiated and virtual experiments need to be run to gain insight for model-based reasoning. The lengthy development and validation process means that agent-based models will be most useful for activities with a multiyear time horizon.

State of the Art

A great deal has been learned about the large number of cognitive biases held by individuals and social groups, social constraints on communication, and the effect of incentives and sanctions on the distribution of goods and activities that are important to consider in models of geoembedded social activity. More than 100 cognitive biases and hundreds of biases in the formation of network ties and the factors that constrain actor access to information have been identified. A large number of factors that influence the existence and strength of relations among actors and under what conditions those relations affect decisions have also been identified. Social system models are becoming increasingly complex and are better representing human behavior (Weinberger, 2011). The realism of models based on cognitive tradition is being improved by imbuing the agents with social cognition. Active areas of research include the following:

- Deviations between actual behavior and normative standards;
- Elicitation of human preferences, beliefs, desires, aspirations, and so on needed to construct realistic models of human behavior;
- Deviations between actor decisions and behaviors when acting independently or in a group;
- Understanding how the social context or structure of the group affects group or societal outcomes; and
- Understanding how to represent uncertainty in an individual model and in combined subsystem models.

The current state of the art in social system models is summarized in Box 4.3. Social system models are complicated to build, are data greedy, and operate at such a transdisciplinary level that underlying theories may not exist. These complexities mean that most models are used for reasoning, rather than forecasting, and that there is little validation and little to no model resuse. On the other hand, building models is facilitated by the availability of toolkits. System-level model toolkits are highly developed, and multiple commercial products (e.g., iThink and Stella for system dynamic models) are in use. Simple models, on which more detailed ones can be built, have been developed. The models can be built to output data directly to various statistical packages for analysis. The methodology has been documented extensively, and many practitioners find that simply creating the model "wiring diagram" is sufficient for explaining how the parts of the system work together. Toolkits for agent-based modeling are not as mature, but they facilitate model development and support integration with the common statistical platforms and network analysis tools used for analysis, validation, and system testing. Common toolkits for the more cognitive models include ACT-R and SOAR, and toolkits for the least cognitive models include Repast, Mason, and NetLogo.

BOX 4.3 State of the Art in Social System Models

• **Space and time scales:** The methodologies used for social system models (e.g., system dynamics models of human behavior) can be used at a wide range of scales. Some models do not consider space; others take into account spatial scales ranging from a small area around a single actor to the entire globe. Likewise, some models do not consider time; others operate at temporal scales ranging from a few nanoseconds (pre-cognitive) to centuries. Spatiotemporal scale tends to be correlated with the number of actors being modeled, particularly for agent-based models. Cognitive agent-based models cover the least space and time, typically those needed for a single task, whereas more general agent-based models and system dynamic models tend to be used for global or multiyear models. Many models do not represent time or space in a "strong" way, and so they provide outcomes only in relative terms (e.g., increasing or decreasing) and present the possible order of actions, but not the time, they occur.

• **Fidelity:** The fidelity of the social system model depends on the phenomena being modeled and the effort invested in the model. The methodologies support any level of fidelity. Because most models are built to demonstrate or explain a general phenomenon, they have relatively low fidelity on all dimensions, except the one(s) of critical concern to the modeler or end user. Most models used for forecasting are partly validated, are tightly tied to some data streams, and were often tuned at least once to some historic event.

• **Accuracy and precision:** The accuracy and precision of these models depends on the underlying theory being modeled and the level of data used to instantiate the model. Most current models are relatively imprecise and tend to be more accurate about general processes and groups than specific events or people.

• **Predictions and scenarios:** Social system models are used not for prediction in the classic sense but rather for suggesting the landscape of possible future scenarios and the relative likelihood of various outcomes. When the scope of the model is narrow and sufficiently multidisciplinary teams are involved in model development, it is usually possible to match the space of possibilities generated by the model with the frequency of events that occur.

• **Uncertainty analysis support:** Few of the modeling methodologies support automatically tracking uncertainty and the propagation of uncertainty through the model. In general, uncertainty is handled by running a large virtual experiment and then examining the distribution of the results. Consequently, analysts talk about the robustness of the results to changing parameters. When the underlying process is not well understood, the process is commonly modeled as random, or as a set of alternative processes which are then compared.

• **Validation and assessment support:** Most social system models violate the assumptions on which validation theory is based (e.g., stationarity of process), and so validation methods worked out for physical systems do not apply to social system models. For example, tuning the inputs and processes to generate outputs that match a historic case generally yields an overtuned model that cannot be used for adaptive actors. The level and type of validation for social system models depends on the purpose of the model. Most social system models never receive more than face validation because they are typically used for reasoning rather than prediction. Validation and assessment are generally easier when the models are built so that the inputs and outputs are in formats that can be used by standard social network tools and statistical packages (e.g., in CSV format).

• **Computational requirements:** System dynamic and agent-based models can generate terabytes of data. When large numbers of virtual experiments are run, distributed processing systems (e.g., cloud computing or a condor cluster) are useful. A laptop is sufficient for small models and experiments. Currently, few models can be turned into Web applications because of both data and processing demands.

• **Data requirements:** It commonly takes more time (sometimes an order of magnitude more time) to collect data to instantiate or test a model than it does to code the models, particularly those used for detailed assessment. Data are often drawn from multiple heterogeneous sources and the diverse data streams need to be fused. The amount of data generated by the models depends on factors such as the number and cognitive fidelity of actors that are modeled, the number of social outcomes tracked, and the number of time periods simulated, which add up to a handful in cognitive agent-based models to hundreds of thousands in population

often exceeds the amount of empirical data that could be collected from the real world on the same topic.

- **Difficulty to develop:** High. Social system models generally take 1 month to 1 year to develop, although many of the findings from traditional work can be applied in a few minutes. Instantiating the models with data and validating them requires at least an equal amount of time.
- **Reuse:** Most models are built once and never reused.
- **Software/code availability:** Multiple modeling toolkits exist for many classes of models. No toolkits have all the findings regarding cognitive biases, social network biases, or game-theoretic considerations built in.
- **Training support:** Textbooks on modeling now exist, but most cover only one type of social science models, such as system dynamics models.
- **Data-to-simulated results:** Many models are paper-only designs and are never actually built, and so no simulated results are generated. A paper model, often referred to as a wiring diagram, is used to illustrate what factors may be influencing others. Such models are commonly used to support reasoning.

How to Make Useful for NGA

For NGA, social system models are most useful for understanding how diverse social and physical subsystems interact to effect new outcomes. The accuracy of these models depends on the correctness of the assumptions, theories, available data, and the process description of the connectors between different subsystems (e.g., a description of how a change in water availability affects health or job availability). Teams comprised of computer scientists, engineers, and mathematicians will not have a sufficient understanding of the social processes, what theories are valid or untested, what data are available, and the current state of the art in modeling the relevant phenomena (Medina and Hepner, 2015). Consequently, the model development team may need to be quite large and include experts in multiple areas, such as geography, history, psychology, organization science, economics, and sociology. NGA has few scientists with a background in these areas. For system-level models, which must be tuned to provide accurate forecasts, data analysts, statisticians, individuals trained in experimental design, and a host of specialists for each critical connector process may also be required. Given these considerations, NGA might be best served by utilizing the expertise and skills of modelers outside of the agency.

For NGA analysts who need to model human behavior, an understanding of the limitations of simple traditional models (e.g., simple game theory and rational actor models) as well as basic training in cognitive biases, network biases, and the role of incentives and sanctions would be invaluable. Such training would help these modelers understand the limits of current knowledge and the areas of uncertainty in modeling geohuman activity, and gain the vocabulary for working with other social system modelers.

Finally, for social system models to be valuable to NGA, they need to be developed within the space of geospatial data. Huge volumes of disparate data are often required to effectively model the system of concern. For example, modeling the potential impact of drought on a population requires data on water levels, water use, laws governing water use, population location, growth and movement, and other data. Given the disparate models and methods used in the different domains of interest to NGA, greatly enhanced data interoperability capabilities will be essential for social system modeling. Particular needs include the ability to search across all NGA data holdings using natural language searches, standardized metadata to enable fast searches and cross correlations, and the ability to bring multiple data streams into a single analysis platform.

NGA-Funded Research and Development Areas

Social system models are useful because they account for the human use of space, the impact of geographic features on sociocultural behavior, and the impact of sociocultural behavior on the built environment and the geography. However, the utility of these models is limited to providing general high-level guidance, supporting model-based reasoning, and demonstrating geosocial behavior in specific settings. The following core research would improve the utility of social system models for NGA:

- Developing new techniques and procedures for increasing the ease of developing, testing, using, and reusing social system models. A particular need is to develop a new theory of validation for social system models.
- Developing an infrastructure for combining models at different levels of resolution and the basic theories, methods, and algorithms for moving between these models at different levels. This simulation testbed needs to support the incorporation, running, and comparison of models built using different paradigms as well as the associated statistical and network analytic tools. A key will be developing a common representation scheme for temporal, spatial, and group features at different levels that can be used with both system dynamic and agent-based models.
- Developing standards for representing spatial information and for collecting and fusing geosocial data.
- Developing data sets at different temporal, spatial, and group levels of granularity that can be used by modelers outside NGA to develop tools and techniques of value to NGA. Such open-access data would allow the broader modeling community to work on geospatial-related social issues.
- Improving understanding of how human behavior is constrained or enabled by the geography of the natural and built environment.

Intelligence questions are time sensitive, and so decreasing the amount of time or personnel required for model development, testing, and use is critical if social system models are to be used more routinely for geospatial intelligence. Key approaches to decrease time and effort include simulation testbeds and methods for making social system models reusable, automating the empirical instantiation of social system models, and conducting sensitivity analysis. Other necessary advances would be enabled by developing a comprehensive representation scheme for geographic factors and shareable geographic data for instantiating the models using these representational forms. Challenge problems for developing social system models using these common representations and shareable data could be beneficial here. Finally, basic research that would improve NGA's modeling capability over the long term includes understanding how geographic factors influence the development of social networks and communications among actors, including covert actors, and how cognitive biases influence the perception of space.

COUPLED PHYSICAL–SOCIAL SYSTEM MODELS

Many geospatial intelligence investigations will involve multiple physical and social system subsystem models and processes that interact with one another. Subsystem models and processes that depend on one another may be coupled to understand the interactions between social and physical systems at different locations. The coupling can either be one way, in which only one model supplies data to the other, or two way, in which both models exchange data (see Appendix A). One-way coupling is simpler to implement and the results are easier to analyze, but two-way coupling tends to yield more realistic results. One advantage of coupling is that simpler and inexpensive models of subsystems (reduced-order models) can be substituted for almost any full model, enabling faster testing and execution of the coupled system. Coupling is increasingly being used to defray the cost of developing large models, to facilitate adaptability and expansion, to leverage the strength of multiple modeling technologies, and to reduce reliance on any one system developer.

Physical and social system models are increasingly being coupled to understand and simulate current and future behaviors and interactions of humans and their environment at various locations and spatiotemporal scales. For example, the megacities intelligence question (see Box 1.2) would likely require coupled physical process–social system models to examine how an urban population may respond to an environmental stressor, such as a heat wave, water shortage, vector-borne disease, or air pollution. Coupled models can focus on individual sectors, such as the transportation system and its effect on the environment,[3] the effect of climate on the energy demands of buildings,[4] or the effects of climate on agriculture (e.g., Figure 2.10) or other ecosystems. More recently, efforts are being made to couple sector-specific models. For example, the PRIMA project links energy-system models, infrastructure models, and regional climate models.[5] Because many problems are spatially heterogeneous, the models being coupled may depend on the specific location of interest.

More complicated interactions and feedbacks among physical and social system processes may be captured in integrated assessment models. Such models combine multiple features of human cultural, religious, and political domains; economic, financial, energy, transportation, and food systems; and the natural world (see Figure 4.4). The Chinese water transfer intelligence question (see Box 1.2) would likely require integrated assessment models to examine the complex interactions among water, agriculture, and energy production and consumption in China. Integrated assessment models can be used to identify vulnerabilities between different systems or regions, and their outputs are frequently used as drivers for other models, such as climate models, conflict models, and regional impacts models.

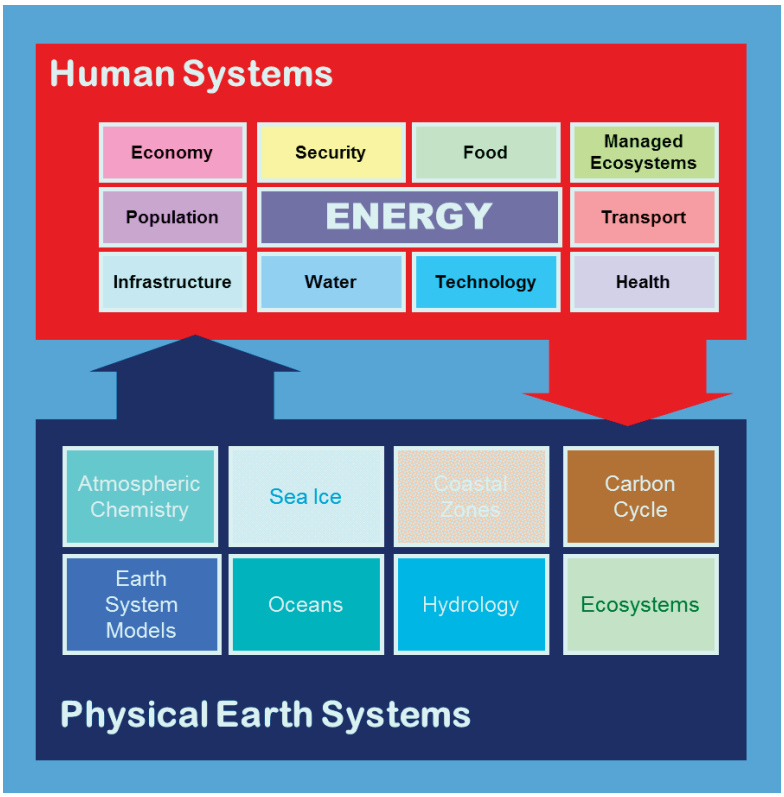


FIGURE 4.4 Schematic of a coupled human–Earth system model. SOURCE: DOE (2009).

[3]For example, see the GREET model, https://greet.es.anl.gov.
[4]See the BEND model, http://prima.pnnl.gov/regional-building-energy-demand.
[5]See http://prima.pnnl.gov.

State of the Art

Coupled models range from fairly basic mathematical forms to complicated network and agent-based formulations that simulate human movement and contact. The state of the art depends on the specific model. A good example of the state of the art is the energy community's integrated assessment models, which simulate large-scale aspects of and trends in technology adoption, economic activity, population growth, and physical system constraints or resources, and are designed to explore different scenarios and outcomes. The state of the art in integrated assessment models is summarized in Box 4.4.

BOX 4.4 State of the Art in Integrated Assessment Models

- **Space and time scales:** From countries to regions (collections of countries) and years to decades. A few models operate on a subnational scale for particular sectors (e.g., agriculture and land use). Specific integrated assessment models and methods are generally designed to simulate specific processes and features within a limited spatial and temporal scale and are rarely valid outside of those scales.
- **Fidelity:** The fidelity of integrated assessment models lies in their representation of interactions between different sectors of the economy. These models often use reduced-form models of an individual sector, for example, representing large groups of actors with a single representative agent.
- **Accuracy and precision:** The accuracy of integrated assessment models is not well quantified.
- **Predictions and scenarios:** The models explore scenarios of how the world might unfold under different conditions rather than provide predictions because of high uncertainty about the future.
- **Uncertainty analysis support:** Sensitivity analysis and multimodel comparison are commonly used to characterize uncertainty, although recent efforts have focused on more formal uncertainty characterization techniques (e.g., Butler et al., 2014).
- **Validation and assessment support:** The community has largely focused on understanding future dynamics and trends, and it only recently began a process of formal model validation over a historical period.
- **Computational requirements:** Computationally inexpensive, a single simulation can often be run on a laptop on the order of hours, although increasing spatial, temporal, and process resolution are increasing the computational expense. Efforts to characterize uncertainty, which may involve hundreds to thousands of simulations, require larger computers and computing clusters. Certain model configurations may also increase computational expense. For example, compared to integrated assessment models with a simple climate component, the integrated Earth System Model (Collins et al., 2015) uses a higher-resolution climate component, resulting in a 10^5-fold increase in computational expense.
- **Data requirements:** Integrated assessment models require large sets of internally consistent data, including information on energy production, consumption, agriculture, land use and land cover, emissions, and the economy. Global data, divided into countries or regions, are necessary. This community is moving toward freely accessible and interoperable data.
- **Difficulty to develop:** High. Most integrated assessment models have been developed by large interdisciplinary teams over years to decades.
- **Reuse:** High. These models are used by multiple researchers for numerous projects and studies.
- **Software/code availability:** Variable. For many integrated assessment models, software and code are limited to those researchers employed by the developers. Some models are now open source.[a] Additionally, some organizations, such as the Global Trade Analysis Project and the Energy Technology Systems Analysis Program, provide software and data used to develop models for a fee.
- **Training support:** Variable. For many integrated assessment models, training support is limited to those researchers employed by the developers. However, some of these models are moving toward a community-based approach, and offer support through annual tutorials, listservs, and online documentation.[a]
- **Data to simulated results:** These models were designed to transform data to simulated results. However, the process by which this occurs differs by model.

[a]See http://www.globalchange.umd.edu/models/gcam.

How to Make Useful for NGA

A number of model-based investigations of interest to NGA will require the use of multiple models. Developing expertise and skills in using different classes of models (e.g., process and empirical), and experience in combining models and comparing their outputs will likely be useful. For NGA analysts who need to work with physical–social system models, it is important to understand that uncertainties associated with the specifications of human behavior often far exceed those associated with the specifications of physical systems. Thus, improving the performance of coupled physical–social system models may best be achieved by improving the social and behavioral aspects of the model.

NGA analysts can learn to run existing coupled models, such as integrated assessment models. However, significant expertise would be required to develop the models further or to interpret the model results. It may be useful for NGA to develop a catalog of scenarios or types of analyses that are scientifically supported and that can be modified for future model-based investigations. In such cases, NGA will need to understand the strengths and weaknesses of integrated assessment models and how to expand the scope of analysis, including designing self-consistent scenarios and developing strategies for analyzing coupled results. Working with developers of integrated assessment models—who have experience coupling models from different disciplines, with different resolutions, and built for different purposes—would likely be helpful.

NGA-Funded Research and Development Areas

The following areas of research could improve the usefulness of coupled physical–social system models to NGA:

- Improved representation of the system components being modeled,
- Methods for formal verification and validation of model results against NGA-relevant benchmarks and test cases, and
- Formal uncertainty quantification techniques.

System components in coupled models are often derived for a specific purpose and it may be necessary to improve their representations for NGA use. For example, because integrated assessment models were developed to analyze long-term climate and climate mitigation, the models typically represent the world in 5- to 15-year time steps, capturing long-term trends and not interannual variability (Krey, 2014). The questions of interest to NGA may require shorter time steps and better representation of short-term phenomena. Formal verification and validation of coupled models is nontrivial because of the level of interaction and communication between the different elements. In integrated assessment models, the heavy dependence on scenario assumptions and boundary conditions makes independent verification difficult. Finally, there are significant uncertainties surrounding the future evolution of human and natural systems. While some uncertainty techniques (e.g., scenario analysis) are widely used, the use of formal uncertainty techniques is somewhat nascent.

INVERSE METHODS

A forward model, such as the physical process or social system models discussed above, requires input parameters to produce outputs (i.e., predictions). These input parameters may describe spatially distributed initial conditions, boundary conditions, and source terms, as well as model coefficients, physical constants, or even model structure. Rarely are all of these input parameters known in advance. Typically system observations (i.e., data recorded at various space-time locations) are required to either estimate these parameters or constrain their

uncertainty so that model-based results are more realistic. The task of inferring these unobserved input parameters from system observations is called solving the inverse problem.

There are two main challenges in solving inverse problems. The first is that the forward model is often computationally demanding to run, making it time consuming to search through the input parameter settings to find values that are consistent with the data. The second is that many inverse problems are ill posed, meaning that many different sets of parameter values may be consistent with the data and their noise. This characteristic of inverse problems, as well as uncertainties in the data and model, imply that uncertainty in the solution is a common feature of inverse problems.

Inverse methods provide systematic frameworks for employing data (i.e., system observations) to reduce uncertainties in those model parameters, bringing the model output closer to the real-world system. As such, they are fundamental to any modeling endeavor. They are most commonly applied to computationally demanding physical process models, such as hydrologic models, where observations of flow and pressure, taken at various spatial locations over time, are used to estimate porosity and permeability of an aquifer. The Chinese water transfer intelligence questions (see Box 1.2) would likely require inverse methods to estimate or constrain key model parameters of a large-scale hydrologic model (e.g., spatially varying permeability, flow rates, and evaporation) to produce plausible predictions of water availability as a function of location throughout China.

Inverse methods may be categorized as (a) simultaneous, where complete model runs are combined with the full set of observations to estimate all unknown parameters simultaneously, or (b) dynamic, where model runs over smaller time intervals are combined sequentially with additional observations to estimate model parameters and produce predictions, conditional on the data observed up to the current time. For the most part, sequential inverse methods, often called data assimilation methods, have been developed to estimate unknown parameters or the state of a dynamic system over time. Dynamic inverse methods include the Kalman filter (Kalman and Bucy, 1961), the extended Kalman filter (Anderson and Moore, 2012), the ensemble Kalman filter (Evensen, 2009), the particle filter (Liu and Chen, 1998), and a wide variety of related approaches. Simultaneous inverse methods have been used to estimate unknown parameters (or states, source terms, etc.) in both transient and steady-state systems (Tarantola, 2005). The various dynamic and static methods offer different strengths and weaknesses regarding computational cost, model fidelity, data completeness, size of the parameter vector, and other factors. Figure 4.5 shows a simultaneous (left) and a dynamic (right) inverse method applied to the same inverse problem.

Inverse methods can be either deterministic or probabilistic. The estimates in Figure 4.4 are probabilistic, showing plausible state reconstructions. The Bayesian paradigm is most commonly used for the probabilistic approach. It seeks to statistically characterize the probability of all sets of parameter values that are consistent with the data, the model, and any prior knowledge of the unknown parameters. In contrast, deterministic inverse methods seek a parameter setting that results in the best match to the data, typically with some penalty on the parameters to render the solution unique (at least locally).

In general, the choice of inverse method depends on the system being modeled (e.g., dynamic or static), what is being estimated (e.g., a few parameters or a million-dimensional state vector), available data (e.g., diversity, accuracy, volume, and velocity), computational considerations, and properties of the forward model (computational demands, processes it captures, and derivative information). It is common to tailor an inverse method to the features of the problem at hand.

State of the Art

Dynamic inverse methods. Simultaneous inverse methods, including deterministic and probabilistic methods, are most commonly applied to physical system models, for example geophysical models. Deterministic approaches to inverse problems are typically formulated as penalized (i.e., "regularized") nonlinear least-squares optimization problems, where the data misfit function (i.e., the squared difference of model predictions with observed data)

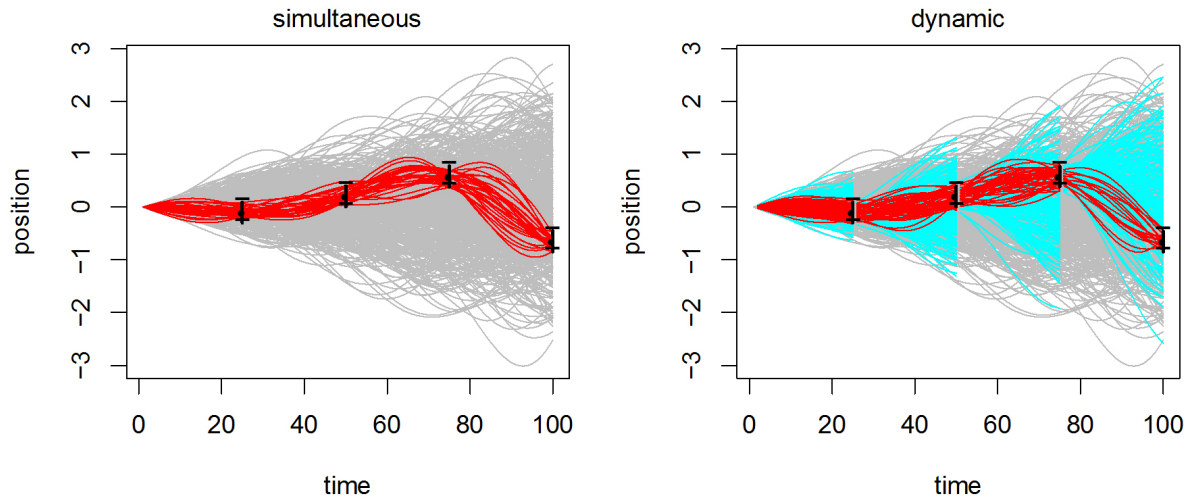


FIGURE 4.5 Comparison of simultaneous (left) and dynamic (right) inverse methods to estimate a one-dimensional state over time. The model allows many plausible evolutions of the state (position) over time (gray lines). At time points t1 = 25, t2 = 50, t3 = 75, and t4 = 100, a measurement is taken, giving the location of the object (with uncertainty) at that time. The simultaneous method uses complete model trajectories along with the entire set of observations to produce plausible trajectories of the object given the model and all of the data (red lines). The dynamic method estimates plausible trajectories over each time interval, combining the model with trajectories produced from previous intervals, the new data point, and the model. Candidate trajectories over the time interval ti to ti+1 are produced by extending plausible trajectories from the previous interval (blue lines). Only trajectories that are compatible with the new data point at time ti+1 (red lines) are kept. This process repeats with each new time interval, giving plausible state estimates given the data up to time ti+1.

forms the least-squares objective function. The state of the art depends on the nature of this objective function and the availability of derivative information from the forward model. When the objective function is nonsmooth, optimization methods that are specialized to noisy functions (e.g., certain direct search methods or simulated annealing) can be used. However, these methods quickly become prohibitively expensive as the parameter dimension grows and the execution of the forward model becomes more computationally expensive (e.g., Earth system models).

When the underlying objective function is smooth and gradients of the function with respect to the model parameters are available, powerful gradient-based numerical optimization methods, often based on Newton's method or its variants, may be employed. These methods can often converge at a cost measured in forward model solutions that is independent of the parameter dimension. If gradients are not available, they can be approximated using finite differencing, generated by automatic differentiation, or obtained by developing an "adjoint" of the forward model (Marchuk, 1995). Of these three, the adjoint method is the most computationally efficient (only a single linearized model solution is required), but it can be difficult to retrofit to legacy codes. Finite differencing is too expensive for large numbers of parameters and often inaccurate for highly nonlinear models. Automatic differentiation is attractive since it requires as input only a forward code, though application to very complex geoscience models is often problematic (but exceptions do exist, such as the Massachusetts Institution of Technology ocean global circulation model). Developing methods for obtaining adjoints of (regularized) nonsmooth problems or legacy codes, as well as automatic differentiation methods that apply to very complex codes, are active areas of research.

Estimating uncertainties in inverse problems typically involves more than just finding the optimal parameter solution. State-of-the-art probabilistic inverse methods use the Bayesian paradigm for statistical modeling, producing a posterior probability distribution for the unknown parameters. This posterior distribution describes the probability of parameters given the data, model, and any prior knowledge on model parameters. The red lines in

Figure 4.5 are draws from the posterior distribution of the state given the data. For nonlinear forward models, the resulting posterior distribution is not in a standard form, and so computing the mean or variance is nontrivial and drives much of the research in this area.

Markov chain Monte Carlo (MCMC) methods are commonly employed to produce samples from the posterior probability distribution. MCMC algorithms range from simple and general (e.g., the Metropolis-Hastings algorithm) to high dimensional and derivative based (e.g., Metropolis-adjusted Langevin algorithms; see Brooks et al., 2011, for a review of methods). The basic approaches are amenable to nonstandard models, even those with discontinuities and regime changes. However, they generally require the forward model calculation to be computed quickly, and they may have difficulty with high-dimensional parameter vectors. Recently developed MCMC methods can leverage derivative information from the forward model, efficiently producing posterior samples, even for high-dimensional parameter vectors. Such methods have proven effective in physical process models that also produce first-, second-, and even third-order derivative information in the course of the forward model run. These highly specific MCMC algorithms are generally developed in concert with the computational forward model and require a substantial amount of expertise in modeling and high-performance computing.

Sequential inverse (data assimilation) methods. The original Kalman filter still influences the research frontier in dynamic inverse methods, yielding an analytic Bayesian solution when the forward model is linear and the errors follow a normal distribution. As with simultaneous inverse methods, the difficulties come with nonlinearities in the forward model, and nonnormality in the errors, both in model evolution and observation. Research is driven by the ever-growing parameter or state space dimensions that arise from increased model resolution, evolving data products (more data, arriving more rapidly), and the computational challenges associated with large-scale models.

Perhaps the most common and successful deterministic data assimilation method is 4D-Var (Ide et al., 1997), assimilating data over a moving four-dimensional space-time observation window. It uses derivative information to produce best estimates of the large-scale parameter (state) vectors, ingesting new data to refine estimates at various intervals, iteratively solving a simultaneous inverse problem over the moving observation window. 4D-Var requires derivative information about the response of the model, and so it is applicable only to systems with smooth input–output maps, such as those found in geophysical systems, such as atmospheric systems. Uncertainty can be estimated using a normal approximation based on the data misfit function used in the 4D-Var procedure.

Ensemble-based methods for dynamic inverse problems (Evensen, 2009; Ott et al., 2004; Tippet et al., 2003) can leverage parallelism to estimate very high-dimensional parameter or state vectors and their uncertainty without the need for derivative information from the forward model. These methods have been used successfully in physical process models (e.g., weather or ocean modeling), although their efficacy for social system models is largely unexplored.

Ensemble-based approaches give approximate draws from the posterior distribution. If the unknown parameter or state vector is small and the forward model can run sufficiently quickly, more exact sequential Monte Carlo methods (Liu and Chen, 1998; Ristic et al., 2004) may be applicable. The dynamic example in Figure 4.5 uses one of these methods. Although limited by the size of the parameter or state vector that can be accommodated, sequential Monte Carlo approaches can handle nonlinear forward models that exhibit rapidly changing behavior and nonnormal error distributions—properties that apply to a number of social system models.

Ongoing research in numerical data assimilation tools for dynamic systems focuses on (a) estimation for cases where the computational cost of the forward problem is high, thus making the use of a large number of iterations or a large ensemble size unrealistic, and (b) the development of more reliable ways to quantify the uncertainty associated with estimates.

The state of the art in inverse methods is summarized in Box 4.5.

BOX 4.5 State of the Art in Inverse Methods

- **Space and time scales:** The relevant scales that inverse methods will inform about will depend on the resolution of both the forward model and data sources being used for the analysis. The space and time scales of the model and data should be comparable.
- **Fidelity:** The fidelity of the results from analyses using inverse methods will depend on the fidelity of the forward model and on the quality of the data being ingested in the analysis. The fidelity of the model is an important consideration in analysis results that describe uncertainty in future scenarios or predictions.
- **Accuracy and precision:** One of the goals of inverse analyses is to produce some characterization of accuracy or precision of the results. The accuracy or precision of the results depends on a number of factors, including the properties of the forward model, the quality of the data, and experience with previous results.
- **Predictions and scenarios:** Deterministic inverse methods produce a single "best estimate" for a prediction; probabilistic inverse methods seek to describe the uncertainty in possible outcomes via error bars, probability distributions, or ensembles of possible outcomes or scenarios.
- **Uncertainty analysis support:** Probabilistic inverse methods quantify the uncertainty in the inverse solution, in both simultaneous and dynamic settings. Deterministic and filtering methods can estimate uncertainty using ensembles or Gaussian approximations.
- **Validation and assessment support:** Validation and model assessment are common elements of inverse methods. How well predictions from inverse methods will fare in new settings depends on the representativeness of both the forward model and the data used for estimation in the new scenario.
- **Computational requirements:** Inverse methods can be applied to small forward models that run quickly on a laptop or to larger models on supercomputers. Some specialization of the inverse methodology is often required to deal with the properties of the forward model. Generally, inverse methods can be hundreds to millions of times more expensive to solve than the corresponding forward model.
- **Data requirements:** Inverse methods require data to estimate model parameters and model structure. The information contained in the data about the parameters dictates how ill posed the inverse problem is, and thus what kind of regularization or prior information should be used, and what additional data sources may improve the results.
- **Difficulty to develop:** Easy (for interfacing simple models with black-box-based inverse methods) to difficult (for developing adjoint-based implementations for derivative-based inverse methods with custom inverse solvers, which requires expertise in optimization theory, adjoint methods, and Bayesian statistics).
- **Reuse:** Inverse methods must be supplied with both method- and problem-specific information, including smoothness penalties, prior information, observations and their uncertainties, and forward models and their uncertainties. Derivative-based methods (as opposed to black box) must be further supplied with derivatives of the misfit between model predictions and observations with respect to the model parameters.
- **Software/code availability:** Open-source and commercial software for solving inverse problems using a variety of methods is widely available. Useful guides for available optimization software include the Optimization Decision Tree[a] and the DAKOTA software.[b] Data assimilation software, using ensemble-based methods, is also available,[c] but requires substantial expertise to implement.
- **Training support:** Textbooks, conferences, and short courses on inverse problems and methodology abound.
- **Data to simulated results:** Given that the forward model and data sources are available, implementation time and effort ranges from hours to interface a simple model to a black box optimization or sampling method to a few years to develop an adjoint/gradient-based method for a complex model with custom inverse solver.

[a]See http://plato.asu.edu/sub/pns.html.

[b]See https://dakota.sandia.gov.

[c]See http://www.image.ucar.edu/DAReS/DART/; https://math.la.asu.edu/~eric/letkf.

How to Make Useful for NGA

A wide variety of inverse methods exists for making model-based results more like the real-world system being investigated. To help narrow down the choices, NGA will first have to determine which models (or families of models) will likely be needed for their investigations, and then determine which types of inverse methods are most relevant, what external collaborations are necessary, and what NGA expertise needs to be developed. Because many NGA investigations are likely to be exploratory, seeking plausible outcomes of complex systems, it is likely that probabilistic, ensemble-based approaches for dynamic systems will be particularly relevant. For these investigations, NGA could do the following:

- Become savvy users of data products and analysis results produced by inverse methods. Results might include satellite data products, weather forecasts, or hydrologic forecasts produced by domain area experts using relevant inversion methods. In these cases, it will be important to understand how to use the data products, how uncertainties were characterized, and product limitations.
- Run or implement inverse methodology developed by outside experts. Inverse methods are available for many mature physical system models (e.g., the Weather Research and Forecasting model[6] can be run in data assimilation mode). Because the inverse methodology is tailored to the specific model, and is often very involved, it will likely be necessary to partner with experts to refine the model-based predictions and understand the strengths and limitations of the results.

NGA Research and Development Needed

The development and use of formal inverse methods for social system models is a research frontier; advances in this direction have the potential to directly benefit NGA's model-based investigations. For example, the megacities intelligence question (see Box 1.2) would likely benefit from the development of inverse methods that use observations to constrain the state of financial, health, transportation, and other urban social systems over time. Specific research and development efforts that could prove beneficial for NGA include the following:

- Developing research partnerships with appropriate collaborators to develop and carry out inverse methods for social and coupled social–technical or social–physical system models and data;
- Facilitating the development of inverse methodology for constraining the plausible states of social system simulation models to be consistent with available data up to the current time; and
- Facilitating the development of inverse methodology to integrate the diverse forms of data that NGA uses and collects (e.g., satellite data, sensor data, geospatial data, and open-source data).

Advancing inverse methodology research for social system models will require partnerships with researchers knowledgeable about inverse methodology and social system models. It will also be important for NGA to work in partnership with these researchers to ensure that advances in this field are geared to NGA applications.

Finally, the diversity of NGA-relevant data may pose research challenges for inverse methods. If NGA wishes to develop inverse methodology that integrates a variety of different data sources, then research may be needed to understand how best to adapt existing dynamic, probabilistic (i.e., data assimilation) tools to their needs.

[6]See wrf-model.org.

SPATIAL STATISTICS, DATA MINING, AND MACHINE LEARNING

Like inverse methods, empirically based models and analysis approaches seek to combine models with data to better understand real-world systems. However, with empirical approaches the emphasis shifts to the data, informing empirical models using techniques from statistics, data mining, and machine learning. The underlying empirical model is typically selected for its simplicity, parsimony, flexibility, ability to handle vast amounts and varieties of data, or ability to effectively exploit the available computational architecture. Figure 4.6 compares an inverse method and an empirical method (Kriging) for estimating the space-time history of a process. In this example, the inverse method used a process model to produce a more accurate reconstruction, which also produced more accurate predictions of the process in the future. However, the inverse method took far more time and computation than the empirical method, which did not use a process model at all. With more data, the accuracy of the empirical method would increase. Moreover, if a suitable process model is not available because the system dynamics are unknown, an empirical model would be able to produce an estimate of the process.

Empirical methods combine system observations, empirical models, and algorithms for estimation and prediction. They are mainly in the purview of the overlapping fields of statistics, data mining, and machine learning. Rather than attempt to disentangle these fields, this section discusses classes of problems, methodology, and applications that are likely to be useful to NGA's modeling endeavor. For example, the megacities intelligence question (see Box 1.2) would likely require empirical methods to discover crime hotspots, increases in unemployment, degradation of neighborhoods, and other urbanization trends that could trigger political, economic, or security problems. The Chinese water transfer questions would likely require models to detect changes in water availability arising from policy decisions on dam building, agriculture, and coal production.

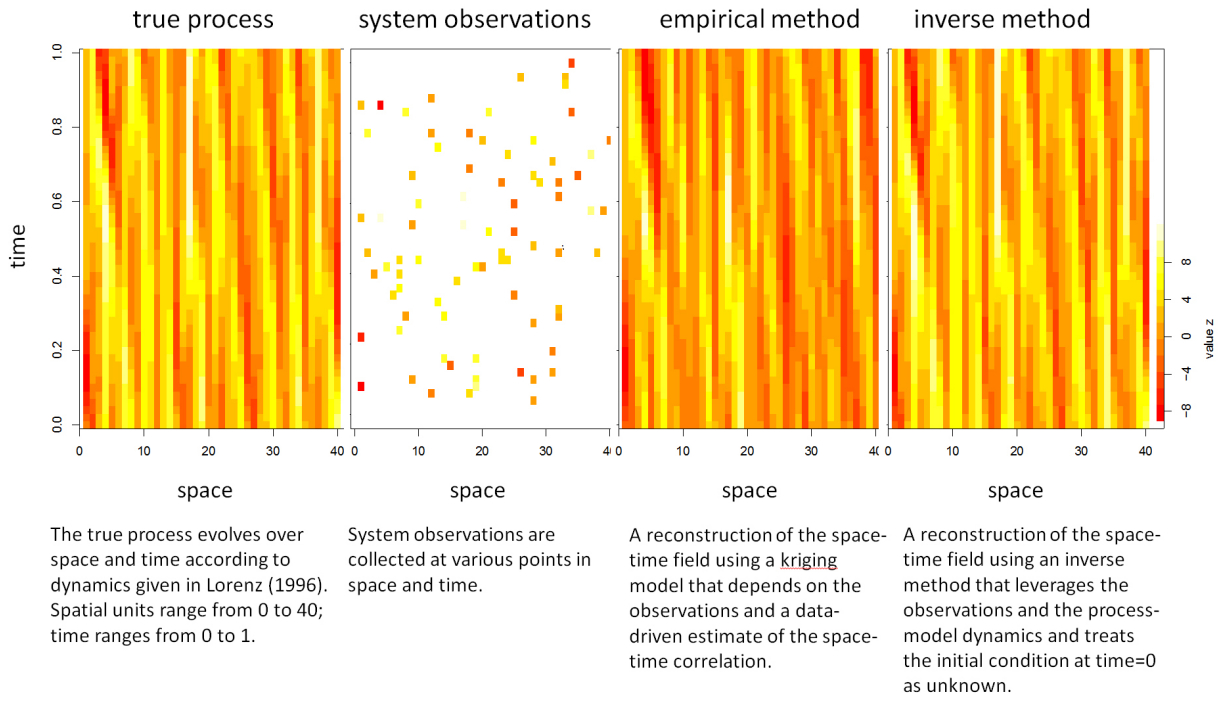


FIGURE 4.6 Comparison of estimates of a synthetic space-time process using an empirical method (combining system observations with an empirical model) and an inverse method (combining systems observations with a physical process model). While the inverse method produces a slightly more accurate reconstruction than the empirical method, it requires far more time and computational effort. SOURCE: Adapted from Higdon (2006).

Empirical models and methods can be divided into supervised and unsupervised learning approaches. Supervised learning approaches entail a wide variety of regression and classification techniques. In such settings, a large collection of observations is used, each containing features X describing the observation, and an outcome response Y, which can be numeric (for regression problems) or categorical (for classification problems). These methods use the observation pairs (X,Y) to estimate a model that predicts the response Y' given a new descriptor vector X'. Hence, the resulting "model" is empirically derived. Examples likely to be of interest to NGA include methods for image classification and methods for prediction using spatially or temporally distributed data. In some cases, the structure of the model is prespecified (e.g., linear regression, classification and regression trees, or neural nets), using the data to estimate model parameters. In other cases (e.g., network models or deep learning methods for image classification), the structure of the underlying empirical model is also estimated from the data. In data-rich settings, it is common to assess the accuracy of such regression and classification models by using a "holdout" data set where the predictor inputs X are made available to the algorithm, but the actual outcome response is withheld.

In unsupervised learning, collections of observations are made, detailing information X for each observation. This information X is used (without any outcome or label variable Y) to identify structure, anomalies, changes, and relationships in the collection of observations. Such approaches may find salient correlations in high-dimensional data and connections between data points that humans would not think to look for. Since unsupervised methods do not require training data, they can be especially useful for exploring large, unstructured data sets. One popular unsupervised learning method is cluster analysis, which is used to find hidden patterns or grouping in data. Other popular methods detect anomalies, hotspots, and changes. Successful applications of unsupervised learning include computer vision, speech recognition, robot control, text mining, and fraud analysis (Alpaydin, 2014; Kou et al., 2004; Sebastiani, 2002; Sung and Poggio, 1998).

For NGA, empirical methods that account for temporal dependence, spatial and space-time dependence, and hierarchical structure are of particular relevance. Methodologies to account for these dependence features and structure include time series (Box et al., 2015; West and Harrison, 1999), space and space-time methods (Banerjee et al., 2014; Cressie and Wikle, 2015; Shekhar et al., 2011, 2015; Zhou et al., 2014), and Bayesian hierarchical models (Gelman and Hill, 2006; Gelman et al., 2013). Specialized methods are required for these situations, because most commonly available software assumes data cases are drawn independently and from a common distribution. Blind application of methods that assume data are independent will typically give biased predictions and can grossly understate the uncertainty. Appropriate empirical methods must use the available data to model the spatial, temporal, and hierarchical dependencies in the system, and they require a familiarity and expertise both in the methods being used and the system being investigated.

State of the Art

Empirical methods and software to specify and implement them have been driven by applications. The most common and generally useful traditional methods (e.g., regression and classification) have long been targets for software development and, hence, have mature implementations that run on a wide variety of computational architectures, ranging from a laptop (e.g., R, sci-py, and MATLAB), to a cluster (e.g., SAS), to data-intensive computing machines running the Hadoop environment (e.g., Salford Systems, Mahout, and Spark MLlib). These mature implementations contain assumptions, such as standard input formats (rectangular data files), common error variance, and independence between data cases.

Mature software is less likely to be available for more complex applications, such as those accounting for specialized model structure (e.g., space-time dependence, model changing with spatial location, and hierarchical relationships in the data) and data features (e.g., multiple, disparate data sources; missing data; large amounts of data; streaming data; outliers; and biases that depend on the data source). For some methods, community-supplied user packages in R, python, and MATLAB may fill the need. For example, many popular software packages implement spatial statistical and spatial data-mining methods, such as spatial summarization, object tracking, trajectory

analysis, hotspot detection, finding spatial outliers, colocations, and location predictions to identify spatial patterns. Methods for spatial summarization include K-Main Routes (Oliver et al., 2014), and methods for hotspot detection include scan statistics and significant ring detection. These approaches are available as R libraries as well as specialized software, such as Crimestat (U.S. Department of Justice) and SaTScan (National Cancer Institute).

On the other hand, the chances of finding appropriate software for a new NGA application are slim. In such cases, methods will have to be developed and tailored for the application at hand. Developing methodology and software is commonly carried out in a high-level programming language, such as R, python, MATLAB, GeoDa (Arizona State University), and STAGE (Joint Warfare Analysis Center). Algorithms for estimation and inference will also have to be developed.

A more recent development in empirical, probabilistic modeling is the emergence of probabilistic programing systems.[7] These systems allow the user to develop a customized model, and the system then automatically produces the computations required to use the data to estimate model parameters and to produce predictions with the appropriate uncertainty. For example, both JAGs and STAN[8] allow the user to develop rather general hierarchical and spatial models, without having to design their own Markov chain Monte Carlo scheme to produce inference results. Such emerging software has the potential to speed up and facilitate the development of stylized empirical methods for nonstandard NGA applications, albeit on standard computing architectures. Software development for spatial methods that leverage modern data-intensive computing remains an open research topic.

Spatiotemporal data management and analytics in modern data-intensive computing architectures have become increasingly important to empirical methods (Eldawy and Mokbel, 2015; Wang et al., 2014). Extensive progress has been made on achieving user-friendly access and interaction with spatiotemporal big data; developing innovative analytics for a variety of geospatial applications, such as traffic prediction (Bast et al., 2014) and disease diffusion prediction (Sadilek et al., 2012); and developing novel data-mining methods for complex trajectories and networks based on massive streaming and sensor data (Tang et al., 2012).

The state of the art in empirical models and methods is summarized in Box 4.6.

How to Make Useful for NGA

To be useful to NGA, the rather large body of methodology available from data mining, statistics, and machine learning will have to be aligned to NGA data sources, applications, and computational resources. This alignment might best be explored and guided using pilot studies, likely in partnership with researchers from academia and industry, so that NGA can better understand strengths and weaknesses of different empirical methods for NGA applications. Focusing on (1) methods and approaches, (2) software and computational infrastructure for their implementation, and (3) data processing and curation could help NGA better exploit available methods in conjunction with computational resources and also better understand what future research directions to pursue. In all of these cases, new and continued partnerships with industry, national laboratories, and university research centers with appropriate cyberinfrastructure and security would help ensure that NGA-relevant systems, tools, and software continue to be developed and refined.

Currently available methods and approaches in statistics, machine learning, and data mining are clearly useful for NGA investigations. However, the standard use cases for which most of these methods were designed may not be directly applicable to geospatial data. Bayesian hierarchical models—particularly ones that link spatially connected data, data at different levels of resolution and aggregation, or disparate data sources—seem particularly useful for NGA. For example, poll predictions of FiveThirtyEight,[9] which combine information from different polls, each with their own spatial coverage and biases, are based on these concepts. Bayesian hierarchical models

[7] See probabilistic-programming.org.
[8] See https://sourceforge.net/projects/mcmc-jags; http://mc-stan.org.
[9] See http://fivethirtyeight.com/tag/2016-presidential-election.

BOX 4.6 State of the Art in Empirical Models and Methods

- **Space and time scales:** These are generally determined by the data on which they are based, as well as the goal of the analysis. Empirical approaches are especially well suited for applications with abundant data.
- **Fidelity:** This is generally determined by the data on which they are based. Fidelity can range from low (using highly aggregated and noisy data) to high (using accurately observed, high-resolution data that are sensitive to all of the subsystem processes and spatial detail involved their generation).
- **Accuracy and precision:** This is a function of the data and what is being estimated or predicted. Generally, more aggregated quantities can be estimated more accurately. For example, yearly national carbon emissions might be estimated to within 10 percent, but weekly emissions from a coal plant may only be estimated to within 100 percent. Systematic biases in data and observations are often inherited by the empirical methods that make use of the data.

- **Predictions and scenarios:** Empirical analyses are typically developed to produce predictions and/or scenarios based on past data. How well empirical models and methods will predict in new settings depends on how representative the data used to train the models are to the new scenario.
- **Uncertainty analysis support:** Most supervised empirical analysis methods come with approaches for quantifying uncertainty in the resulting predictions and other inferences. In spatial and space-time settings, estimation of uncertainties is more challenging because autocorrelation and spatial dependence must be accounted for. The quality of estimated uncertainties is often assessed by comparison with past data, when available.
- **Validation and assessment support:** Empirical approaches are data driven; hence the processes of model validation and assessment are typically part of the analysis. Such assessments typically assume that past performance of the model reflects future performance. Model validation and assessment for extrapolative settings remains an open research topic.
- **Computational requirements:** Some approaches can be developed and fit using a laptop, although demand for compute- and/or data-intensive resources is increasing. In some cases, the large volumes of data are preprocessed using large-scale computational resources, prior to more involved empirical analyses. A current challenge is adapting spatial methods to modern data-intensive computing architectures to handle large-scale geospatial data.
- **Data requirements:** Empirical models require data to estimate model parameters and model structure. The properties of data (e.g., size, type, cadence, resolution, and accuracy) influence what empirical modeling methods are likely to be useful for a given scenario. Many models and methods must be adapted to use large amounts of data and data-intensive computing resources.
- **Difficulty to develop:** Many empirical methods are available in high-level software languages, such as R and python, and so development is largely restricted to preparing data. Special considerations, such as large data volume or specialized spatial or temporal structure, typically require stylized models and estimation procedures, and additional time and expertise to develop.
- **Reuse:** Many approaches can be reused in new settings, although the models will have to be retrained using data from the new setting. In addition, a fair bit of effort is often required to clean and preprocess data for empirical models and methods.
- **Software/code availability:** Open-source and licensed software is available for fitting and developing empirical models. More specialized methods require in-house development.
- **Training support:** Textbooks on data mining, statistics, and machine learning are available; and conferences and short courses are plentiful.
- **Data to simulated results:** Once models have been fit to the data, simulation of outcomes or data is generally straightforward. Assessing the quality and accuracy of the simulated results is more involved.

incorporating output from social or physical system models are less common and require additional development. Other useful methods include (a) clustering and other unsupervised and deep learning methods for finding structure in large volumes of data that are too much to analyze with a human in the loop and (b) methods for detecting change footprints and spatial hotspots and anomalies.

The software and computational infrastructure available for carrying out such empirical analyses may not align well with NGA needs. It is important for NGA to understand the limitations of currently available software (e.g., data volume that can be handled and computational time required for analyses) as well as new software for use on data-intensive computing architectures (e.g., cyberGIS, Hadoop, PY-SPARK, and R-SPARK). Pilot projects and case studies would allow NGA to evaluate software implementations of empirical methodology and determine their applicability to NGA applications and data. Software for spatial statistics and spatial data mining methods (e.g., hotspot detection via SatScan or CrimeStat) could be tested with U.S. Department of Defense and NGA data sets and use cases.

New methodology that leverages data-intensive computational architecture is just coming on line. For problems exhibiting high spatial complexity, emerging tools (e.g., GIS Tools for Hadoop, and SpatialHadoop) and the development of cyberGIS capabilities for exploiting high-performance and cloud computing can facilitate the integration of rich spatiotemporal data, analytics, models, and visualization. These approaches have proven useful for knowledge discovery in a number of domains, including geospatial intelligence, agriculture, hydrology and water resources, coupled physical–social systems, emergency management, geophysics, econometrics, and urban studies (Anselin and Rey, 2012; Wang and Zhu, 2008).

Data processing and curating approaches will be needed to align NGA's geospatial data and empirical modeling efforts. It will be necessary to experiment with approaches for data selection, cleaning, wrangling, and other preprocessing required to prepare data for these methods. These preprocessing steps, often demanding in their own right, can have a substantial impact on the results. It may also be necessary to develop data curation technology or methods to facilitate searching, handling, extracting, and using NGA data.

NGA-Funded Research and Development Areas

Research needs will likely become more apparent as the universe of potential NGA applications is aligned with available methodology and technology. The special nature and variety of NGA's data sources (e.g., satellite, sensor, traditional and nontraditional geospatially indexed, social media, open source and classified, and statistical) will drive research needs in supervised and unsupervised learning approaches. Many challenging problems that NGA faces require in-depth integration of diverse domain-specific and geospatial models with spatiotemporal data management and analytics. How to achieve this integration across various computational and spatial scales (Cao et al., 2015; Das Sarma et al., 2012) requires substantial research. Specific research directions that could serve NGA's future needs include the following:

- Developing methodology to combine diverse, disparate data for inference and decision making, including methodology to combine predictions or results of different approaches (e.g., expert judgment, neural net-based classifier);
- Developing capabilities for accessing and formatting disparate data in ways that enable analysis products to be generated quickly for decision making;
- Advancing unsupervised learning approaches for NGA-specific data and applications;
- Developing new parallel formulations of spatial database management systems (e.g., SQL/OGIS standards), spatial statistics, and spatial data-mining tasks on current platforms (e.g., GPU, clusters, HDFS, MapReduce, SPARK) and on future advanced computing and cyberGIS platforms for modeling work;
- Developing a testbed of common geospatial intelligence analysis tasks (e.g., detection and anticipation of

spatial anomalies, hotspots, and patterns of life), proxy data sets (e.g., trajectories, maps, and satellite imagery) of different sizes, and computational intensity metrics (e.g., efficiency and scalability); and

- Developing novel spatial and spatiotemporal methods that can take direct advantage of advanced data-intensive computational resources, such as formulating spatial processes as parallel processes that can be mapped naturally to parallel computing architecture (NASEM, 2016b; Tang and Wang, 2009).

The research directions given above should help NGA identify promising empirical analysis methods and then adapt them to NGA investigations. These adaptations may focus on extending methodological approaches designed without considering spatial or temporal autocorrelations. They also leverage new and rapidly evolving data-intensive and high-performance computing capabilities for two key tasks: (1) carrying out scalable geospatial modeling, based on advanced spatial and space-time analyses, and (2) integrating large-scale geospatial databases.

SPATIAL NETWORK ANALYSIS

Human sociocultural activity and the results of that activity are constrained by network dependencies. For example, individuals are influenced by those with whom they interact, organizations are constrained by their transaction networks, and countries are constrained by their alliances. Network analysis models that take into account these dependencies are a core technology for cultural geographic assessment and media analytics, a current area of emphasis for NGA. In addition, NGA's megacities intelligence question (see Box 1.2) would likely require social network models to analyze agreements and trade networks among organizations that could help or hinder a response to natural, economic, or political disasters (e.g., Zhong et al., 2014). The Chinese water transfer questions would likely require social media analytics to determine which neighborhoods and which groups are likely to strongly resist migration.

Models and methods that take network dependencies into account are referred to as social network analysis, social media analytics, network science, link analysis, dynamic network analysis, or high-dimensional network analysis. They are a flexible class of graphical and statistical models that explain behavior in terms of the relations among entities. Simple social network analysis models focus on people and how they are related, such as through friendship, financial, or collegial ties. Dynamic network analysis and other high-dimensional variants include many classes of entities (e.g., people, organizations, ideas, resources, and locations) and many classes of ties (e.g., communicates with, borders, resides at). Network analysis models support reasoning at multiple levels (e.g., among people, organizations, or countries), and communications assessments for diverse media (e.g., social media analytics). High-dimensional and dynamic network analysis models can also be used to assess (a) change over time or space, (b) the spatiotemporal constraints on human activity within a specific environment, or (c) the effect of a change in constraints across different geographic regions (e.g., Kas et al., 2012; Medina and Hepner, 2011; Van Holt et al., 2012). Finally, social network tools can be used to support course-of-action assessment and, in certain cases, prediction, both of which are important for geospatial intelligence applications. For example, social influence models can be used to predict things like adoption of technology, formation of opinions, and change in belief. However, these models only take spatial factors into account by altering the strength of ties between actors to reflect distance, and by developing different social influence models for different regions.

Increasingly network modeling is being combined with other spatial statistical and machine learning techniques to help analysts reason using content with large high-dimensional network data at different temporal, spatial, and group levels of resolution. Use of machine learning and language technology adds the capability of a learning or training process. The emergence of spatial network analysis (e.g., Figure 4.7), which combines spatial and network reasoning, has led to new techniques, including a method for assessing information loss across different levels of resolution (Olson and Carley, 2008), identification of actor tweet location using social network information, and

techniques for moving between trail data (e.g., who was where when) and networks (Davis et al., 2008; Merrill et al., 2015).

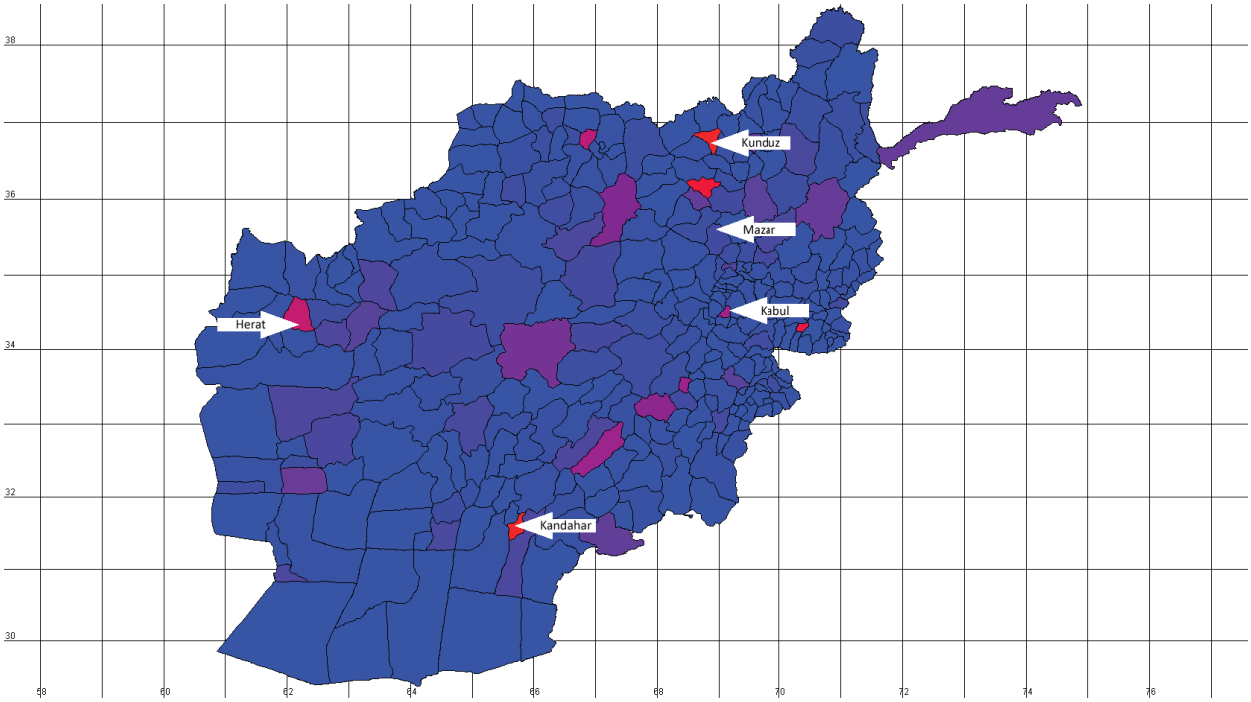


FIGURE 4.7 Heat map of Afghanistan in which each sector is colored by the average degree of centrality of the actors of interest who were geotagged within this region. The bright red sectors are those where there were more actors who were more connected to other actors. SOURCE: Based on an unpublished study of Afghanistan using data drawn from open-source data by Kathleen M. Carley and members of the CASOS team, Carnegie Mellon University.

State of the Art

In the past, traditional network models were focused on only one or two types of nodes (e.g., just people), a small number of nodes (i.e., less than 30), and a single time period. Such models, and the tools that support them, are the most common use of network modeling today. They can be useful to NGA for examining and reasoning about regional heterogeneity (e.g., models showing which organizations work together in different cities to support disaster response). Modern network analysis models, however, are more comprehensive and may have many types of nodes, have large numbers of nodes, and cover many time periods. These modern models and associated tools are more useful, because they can represent and use a wider range of geospatial data.

Network analysis models can be developed rapidly and are often used to assess open-source data (including social media), human intelligence, social intelligence, and survey and simulation results. Some data are also gathered from subject-matter experts (e.g., which leader is hostile to which). Data from multiple sources can be merged as long as the node IDs are matched. This is important because many sources of network data may have few if any spatial data. For example, few surveys track the location of the actors, only a small fraction of social media data are geotagged, and spatial data are often obscured (e.g., locations are made up or indeterminate) or intentionally hidden (e.g., cyberattacks spoof the IPs). Thus, data cleaning, link inference, and merger of multiple data sources is often needed to infer more of the location information.

While it is well recognized that social networks are spatially embedded, relatively little is known about how

that embedding constrains and enables behavior (e.g., Barthélemy, 2011). Individuals are more likely to interact with those nearby, even if they have access to social media and the Internet. Population density within a region affects the sense of anomie, and interactions tend to atrophy as distance increases. However, there is no comprehensive source of all such findings.

Active areas of network modeling research of particular relevant to NGA include the following:

- High-dimensional and dynamic network analysis, with particular attention to n-mode clustering techniques;
- New scalable routines and or approximation techniques for large dense networks, particularly when they take into account spatial relations;
- Linking social networks with other networks and/or node attributes (e.g., spatial network analysis, where the nodes have connections to each other and positions on maps); and
- Assessing the robustness of measures for filling in missing data, and the inference of missing links (e.g., using temporal and spatial dependencies to infer social interactions, and using social interactions to infer geographic location).

Most metrics for assessing network models and community detection algorithms have been optimized for large, scarce data sets, and are quite scalable. These are useful when, for example, assessing the communication networks in social media in different countries. Metrics and community detection algorithms for dense networks, such as shared hashtag networks and high-dimensional networks, are still in their infancy and tend to scale poorly. In addition, specialized visualization tools and metrics for large-scale social media data, particularly Twitter, as well as new techniques for creating networks from streaming data are starting to appear (e.g., Hannigan et al., 2013).

The vast majority of network metrics and algorithms are focused on finding critical nodes, critical links, and groups; characterizing the topology; comparing and contrasting networks; and identifying change over time or predicting one network from other information or prior networks. However, data are not always sufficient to carry out these tasks. For example, metrics designed to identify critical nodes are sensitive to missing data, and in some situations, as little as 25 percent missing data has led to an 80 percent reduction in accuracy (Borgatti et al., 2006; Frantz et al., 2009). Few tools provide support for uncertainty assessment. Data collection strategies and the technologies that generate data can also create biases that influence what can be identified using network techniques. For example, snowball sampling from a single source (as might be done with cell phone data) can overemphasize the importance of the phone owners. The importance of an original tweeter can be overemphasized, because Twitter links both the tweet and any retweets to the original tweeter.

Spatial network measures and techniques are beginning to emerge. The capability to visualize networks on maps is well established (e.g., Olson and Carley, 2009) and toolkits are now available in three packages (Palantir, ArcGIS, and ORA). Many social media and cyberattack collection tools also show some network data on maps, but those tools are proprietary and are not used for analysis. Measures and techniques designed for spatial networks include "spatial" centrality measures, guidance for using links (e.g., when to use link inversion) when representing spatial distance, and network-based measures for assessing autocorrelation (variants of Moran's I and Geary's C). New bipartite spectral algorithms for clustering spatial and nonspatial data simultaneously are in the experimental stage. Tools for engaging in spatial network analysis are often built from scratch by researchers. Some tools are available in ORA and R, but ArcGIS offers only the most primitive of basic network metrics.

Analysts can now use suites of interoperable tools in a data-to-model workflow to conduct rapid ethnographic analyses and understand the lay of the land. Figure 4.8 shows a typical workflow in which the technologies used are interoperable, with the output from one modeling or analysis tool feeding into the next. An interoperable tool suite, rather than an integrated system, is used because of the rapid rate of change in the application program interfaces, the evolution of language technology, and the need to interface with data and reporting tools on different platforms or networks. Many of these workflows are instantiated in frameworks like Ozone or, more commonly, through the use of python scripts.

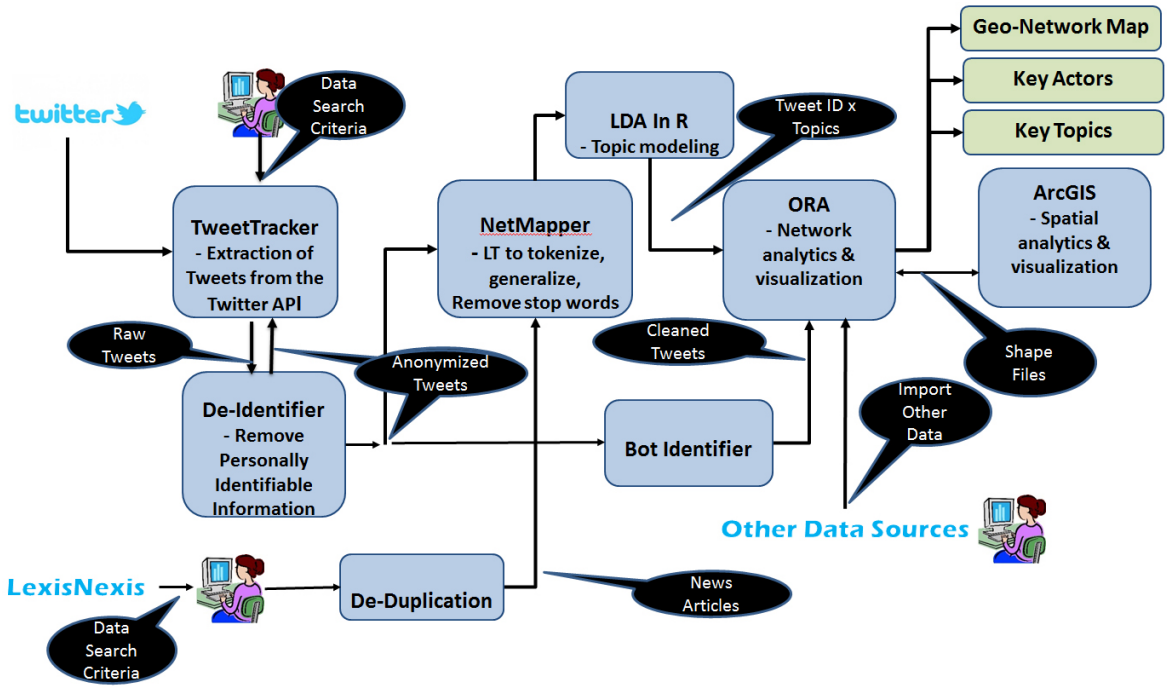


FIGURE 4.8 Illustrative tool chain for network modeling. Data are collected from various sources, cleaned to remove nonrelevant information, and then information is extracted for the analysis.

In this workflow, data are collected from various sources. For example, special technologies may be used to download news articles and social media data. The key challenge is designing the search strategy so that the information of interest is part of the extracted data stream, the data stream does not contain such a high volume of irrelevant information that it overwhelms the analysis tools, and the timeframe for the collected data is appropriate. Special training and experience are generally needed to develop such queries. Next, the data are cleaned to remove data that are not relevant (e.g., tweets by bots) and often fused. Specialized tools are used for bot detection, deduplication, and topic identification. Finally, information needed for the analysis is extracted from the raw text data and metadata. Advanced language technology models are used to tokenize the raw text and then remove unique or low information concepts, and identify bigrams, actors, and locations. Sentiment analyzers may also be run and the sentiment attached to the tweet or the tweeter. The use of parallel processing and semiautomated tools such as NetMapper and ORA have reduced the time it takes to carry out these steps to as little as two weeks (Carley et al., 2012a).

The current state of the art in spatial network analysis is summarized in Box 4.7.

How to Make Useful for NGA

Network analysis models are likely to be useful when NGA is concerned with how patterns of relations affect behavior, whether that behavior is at the individual, group, organization, or state level. These models are also useful for characterizing and reasoning about various types of networks, such as transportation, communication, and infrastructure networks. Because the basic methods are already so developed, NGA could quickly increase

BOX 4.7 State of the Art in Spatial Network Analysis

- **Space and time scales:** Most network models do not consider space or time, although spatial and network regression techniques are used with temporal data. A few tools provide support for spatial and temporal networks. In these cases, the scale of the data determines the scales used in the models because the methodologies can support any space or time scale.
- **Fidelity:** Network models are empirically driven and typically have reasonably high fidelity. However, the model fidelity depends on the data. For an investigation focused on interaction, for example, the model fidelity is higher if phone call data are used and lower if data on co-presence at events are used.
- **Accuracy and precision:** These depend on the data used to build the model, the amount of missing data, and the level of inference needed to place actors or resources in specific situations.
- **Predictions and scenarios:** Network models are often part of a scenario description. What-if analysis using comparative statistics, where things are moved, locations barred, or nodes added or removed are used for prediction. Temporal trends in spatial or network data, and certain metrics in network models (e.g., cognitive demand for predicting emergent leaders) are also used for prediction.
- **Uncertainty analysis support:** This is difficult, with existing tools providing limited technical support.
- **Validation and assessment support:** All existing metrics in commercial off-the-shelf software have been validated. A few new metrics in open-source software have not been validated. It may be possible to use bootstrapping techniques to assess the robustness of model results against missing data.
- **Data requirements:** Network data are increasingly available from social and printed media, sensors, and surveys. Only a few tools (e.g., ORA, some of the i-graph routines in R) support analysis of networks with more than 100,000 nodes and their spatial relations. Many of the more interesting measures cannot be parallelized to run in a MapReduce framework. Approximation and incremental techniques are being developed. Processing the large data is often made possible by using the visualization cards in desktop machines. As the number of nodes (network size), the fraction of possible edges that are nonzero (density), and the number of networks increase, memory limits may be reached and disk processing may be too slow. In those cases, analysis is done in a distributed fashion or by using scripting techniques in batch mode or with special hardware. For most modern machines, these limits tend to be reached when there are more than 10^7 nodes, or in large networks with densities greater than 0.4, or for more than 1,000 large networks.
- **Difficulty to develop:** Low. Network models can be developed in minutes to hours. Spatial network models take longer. Increasingly, text-mining techniques are used to build network models from open-source texts.
- **Reuse:** Modeling technology has unlimited reuse. Specific instantiated models are often reused.
- **Software/code availability:** Multiple commercial off-the-shelf toolkits are available and some open-source tools also exist. A few tools have limited support for both network and spatial data.
- **Training support:** Textbooks are available on developing and accessing network models, collecting data to support them, and interpreting visualizations. Many conferences, special short courses, and train-the-trainer programs are run on a regular basis.
- **Data-to-simulated results:** Many network tools are designed to work in data-to-simulation workflows. Simple diffusion or robustness simulations are part of some network modeling tools.

its modeling capability in network analysis (a) by having staff with strong spatial network competence and (b) by obtaining spatial network data to use for training and for developing more advanced methods.

Few analysts have strong skills in network or spatial network analysis. If they are trained in GIS capabilities, they likely have little understanding of the strengths and limitations of network methods, and may use only the most basic measures and make mistakes (e.g., applying measures to multimode data without separating the modes). If they are trained in social networks or network science, they likely have little understanding of the strengths, limitations, and advances in spatial analytics. In this case, they may limit themselves to creating and comparing different networks for different regions and may not consider autocorrelation issues.

Consequently, developing expertise and skills in network analysis models, particularly in dynamic network

analysis and high-dimensional network analytics, would likely be helpful. NGA modelers with this expertise could help integrate information from different sources, assess social media, and examine issues of resiliency, power, and influence, which are valuable for assessing regional differences. Developing expertise and skills in spatial network analysis would support assessments of the impact of logistical constraints, spatial environment, and spatial autocorrelation on social behavior. Skills in using existing tools to place networks on maps, assess the spatial distribution of ties, and identify the geographic span of key network members would also be valuable.

In addition to hiring staff with the necessary knowledge and skills, NGA could develop in-house expertise. Network and spatial network analytic and visualization capabilities could be taught at both an introductory and a more advanced level in the NGA College. In addition, NGA could send more personnel to university-based programs that train the trainer in network analytics for social media, high-dimensional network data, and spatial network analysis. Most training programs conducted as executive education or conference seminars are focused on traditional network analysis and do not cover spatial or temporal issues. Thus, NGA analysts may need to go both to sessions on traditional network analytics and to sessions on spatiotemporal network analytics.

Although simple spatial network data sets (e.g., air traffic flow) are available through Carnegie Mellon University, developing more relevant data sets would be invaluable for training. In addition, applications of spatial network models could be improved or refined using a number of different methods for analyzing and fusing data. For NGA, tools to search and fuse classified and open-source data would increase the speed with which these modeling technologies could be applied in new settings. For example, a simple but valuable technology would add spatial coordinates based on place names in a format that could be imported into network tools.

NGA-Funded Research and Development Areas

Network analysis models would be useful for a wide range of NGA applications. However, the current techniques have three key limitations. First, the availability of spatiotemporal network data is growing, but most methods cannot cope with space, time, and groups simultaneously. Second, many of the existing technologies support description more than prediction. Third, spatially tagged data are sparse, and even when the data exist, theories about geographic constraints on networks and network constraints on geography are fairly weak. NGA-funded research and development that could help overcome these limitations include the following:

- Developing joint spatiotemporal-network techniques for assessing and visualizing the relations among the entities and activities of interest in high-dimensional data (e.g., autocorrelation visualization techniques);
- Developing joint spatial network models to predict the spread of information, technology, and activities; the engagement of entities in activities of interest in regions of interest, given existing social networks; and the development of networks and activities of interest, given geographic constraints such as transportation and communication barriers;
- Developing spatiotemporal-network-based sentiment-mining techniques for informing predictive spatial network models;
- Developing data sets with multiple levels of spatial and social network data for use in developing theories and testing metrics;
- Commissioning review papers that build a compendium of findings about the relation between spatial factors and network factors; and
- Improving training in joint spatial network analytics and visualization.

Joint spatiotemporal-network techniques are needed because spatial statistical models and methods are limited to isotropic Euclidean spaces, which are inadequate both for spatiotemporal data (e.g., trajectories) and for spatial network data (e.g., crime reports in urban areas). In addition, network clustering and prediction techniques rarely take spatial or temporal distances into account. Specific research areas include (a) accounting for spatial

autocorrelation or coverage when assessing centrality (e.g., actors that have more ties in more locations may be more critical than those with more ties in one location), (b) assessing when network structures migrate between spatial regions (e.g., as can happen when a gang moves to a new city but retains their network structure), and (c) developing techniques for assessing networks where the ties represent physical distance and for using spatiotemporal information to help identify covert networks. A good starting point would be to survey past uses of network analysis models that address dynamic spatial situations, such as air traffic monitoring or hostility networks.

Spatial network models that predict the spread of ideas, beliefs, activities, and technologies are needed because existing diffusion models in spatial analytics and network analytics are not integrateable, are unlikely to yield the same predictions, and have insufficient predictive capability. Research on spatial-network diffusion models, particularly reusable models, would improve forecasts of changes in human geography.

Extracting sentiment for large geolocated groups will be challenging. Sentiment miners currently characterize the sentiment in a document by the number of positive and negative words, usually English. However, the results can be misleading, because they do not provide guidance on what the sentiment is directed toward, and they do not account for negation or sarcasm. NGA would gain the greatest value by developing new sentiment assessment techniques that refine the sentiment based on the geographical and network features of the community, such as the confluence of information from time zone, location, local events, language, structure of local groups, and tendency to self-identify in networks through linguistic cues.

SUMMARY AND CONCLUSIONS

NGA will need a wide variety of models and analytical methods to improve its geospatial intelligence capabilities. For example, models and methods needed to answer the megacities intelligence question provided by NGA (*How will worldwide urbanization trends affect regional political, economic, and security environments?*) include the following:

- Physical process models of environmental changes that could stress urban populations, such as sea-level rise and increases in summer temperatures;
- Social system models to determine what changes in social systems (e.g., financial, cultural, ethnic, religious, and health) may trigger political, economic, or security problems across different geoterrains;
- Coupled physical process–social system models to examine how an urban population may respond to an environmental stressor, such as a heat wave, water shortage, vector-borne disease, or air pollution;
- Inverse methods that use observations to constrain the state of financial, health, transportation, and other urban social systems over time;
- Empirical methods to discover crime hotspots, increases in unemployment, degradation of neighborhoods, and other urbanization trends that could trigger political, economic, or security problems; and
- Social network models to analyze agreements and trade networks among organizations that could help or hinder a response to natural, economic, or political disasters.

Models and methods needed to answer the Chinese water transfer intelligence questions (*How do agriculture and energy production and consumption change over time? How and where will populations, including rural communities, shift?*) include the following:

- A large-scale physical process model of the hydrologic system in China to predict surface flow, subsurface flow, and abundance of water under different water diversion scenarios;
- Inverse methods to estimate or constrain key model parameters of a large-scale hydrologic model (e.g., spatially varying permeability, flow rates, and evaporation) to produce plausible predictions of water availability as a function of location throughout China;

- Integrated assessment models to examine the complex interactions among water, agriculture, and energy production and consumption in China;
- Empirical models to detect changes in water availability arising from policy decisions on dam building, agriculture, and coal production;
- Social system model scenarios of how affected populations are likely to respond to dam construction and involuntary migration; and
- Social media analytics to determine which neighborhoods and which groups are likely to strongly resist migration.

Steps NGA can take to develop the sophisticated modeling and analysis capability needed to address these types of questions are summarized below.

Develop or Adapt Models or Methods at NGA

Data-driven models and analysis methods are amenable to near-term development, because NGA analysts already have some relevant knowledge and experience, the methodology is established, and software and training support are available. In particular, NGA's experience with spatial and temporal analysis provides a foundation for developing or adapting spatial statistics, data mining, and machine learning methods. Methods that are especially promising for NGA include (a) Bayesian hierarchical models that link spatially connected data, data at different levels of resolution and aggregation, or disparate data sources; (b) clustering and other unsupervised and deep learning methods for finding structure in large volumes of data that are too much to analyze with a human in the loop; and (c) methods for detecting change footprints and spatial hotspots and anomalies. In addition, NGA's growing emphasis in human geography provides a foundation for developing network analysis models to examine how patterns of relations affect behavior.

For both types of analyses, the basic methods are well established, and software and user support (e.g., textbooks, conferences, and special short courses) are readily available. However, some additional development and training are required to adapt these methods for geospatial data and NGA use cases. In addition, software and algorithms for data-intensive computing will likely have to be developed for spatial statistics and spatial data-mining methods. Training in network or spatial network analysis could be offered at the NGA College or obtained from university-based programs.

Collaborate with Outside Experts

NGA will need partners to help develop, adapt, and use more sophisticated models and methods (e.g., process models, coupled models, agent-based models, inverse methods, and spatial network models) as well as geospatial models that are not well supported by cutting-edge computational infrastructure. A substantial part of any group's capability in sophisticated modeling is learned through partnerships, apprenticeships, and collaborations. Such collaboration could take many forms, including being a partner in the team developing a model or extending its use to other applications, a user of a team's model or method, or a user of the resulting data products. Regardless, NGA will need to identify domain experts who can design models or scenarios relevant to NGA, run the model, interpret the results, or help NGA find useful existing model output. To use these models or model results effectively, NGA will need to understand their strengths and limitations for the geospatial intelligence task at hand.

Finding partners for NGA modeling efforts will not be trivial because of the classified nature of the work, the wide and changing variety of experts needed, and the need to nurture long-term relationships. Models of complex systems are typically developed by multidisciplinary teams with in-depth knowledge and experience in the scientific disciplines and computational capabilities relevant for the task at hand. However, bringing together diverse experts, who would learn from each other in the context of NGA's priorities, could contribute to major

breakthroughs in NGA-relevant problems. Major research universities, as well as organizations for which NGA has established relationships (e.g., defense and intelligence agencies, national laboratories, private-sector contractors, and NGA centers of academic excellence in geospatial science) may be a starting point for finding experts and modeling teams for NGA modeling efforts.

Fund Research and Development

Investments in research and development could strengthen NGA's modeling capabilities in the years ahead. Where to focus these investments depends on what models and analysis methods are proving most useful for geospatial intelligence. Potential research areas are summarized below.

Extending the use of models to NGA-relevant situations. NGA investigations will likely make new demands on models, using them in settings for which they have not been originally designed, or at least not thoroughly tested. Examples include precise, near-real-time wind, wave, or weather predictions to support troop deployment, disaster relief, or dispersion and damage estimates from the release of hazardous materials in urban environments. In addition, such models may need to be combined with social system models to help decision makers prepare for social unrest, disruption, or migration. Substantial research is required to adapt physical process models, social system models, and combined physical–social system models to deal more reliably with these less common settings.

Improving understanding of human behavior. Social system models are only beginning to surpass expert judgment. Advancing their development, and the development of combined physical–social system models, requires fundamental research to improve understanding of human behavior. Promising areas of research include studies aimed at understanding how human behavior is constrained or enabled by the geography of the natural and built environment, including how geographic factors influence the development of social networks and communications among actors, and how cognitive biases influence the perception of space.

Speeding model development, testing, and run time. Intelligence questions are often time sensitive, and so research advances that speed up model development, testing, or run time could prove beneficial to NGA. Model development could be sped up through research aimed at facilitating the combination of existing subsystem models for NGA investigations. Model development and run time could be decreased through research and development of accurate, reduced-order models or emulators that effectively reproduce results of computationally intensive models. Developing simulation testbeds could aid all of these efforts and also facilitate assessments of model accuracy and speed.

Methodological research and development tailored to NGA-relevant models. The models developed or adapted for NGA purposes will have to be accompanied by customized methods that combine these models with data. Methodology for inversion, exploration of plausible outcomes or scenarios, quantification of prediction uncertainties, and model assessment will be required to bring model-based results more in line with available measurements. Such methodology is particularly needed for social system and physical–social system models. Possible directions in this area include development of inverse methods for constraining the plausible states of social system models to be consistent with data, and development of methods for formal verification and validation of model results against NGA-relevant benchmarks and test cases. In addition, research on how to adapt existing inverse methods to integrate the diverse forms of data that NGA collects and uses (e.g., satellite, sensor, geospatial, and open source) would be beneficial for all types of models.

Methodological research and development tailored to NGA data sources and needs. Research could facilitate the development of empirical methodology that is tailored to the data used by NGA. Examples include developing

methods to combine disparate data or results from different approaches and more accurately represent their uncertainty to support inference and decision making, and methods to cope with data that have spatial, temporal, and network components (e.g., to assess the activities of a terrorist cell over time). A related need is for sentiment-mining techniques that characterize the sentiment in a document based on the geographical and network features of the community, such as location, local events, language, structure of local groups, and tendency to self-identify in networks. Research is also needed to develop spatial and spatiotemporal methods that use advanced data-intensive computational resources, such as formulating spatial processes as parallel processes that can be mapped naturally to parallel computing architecture.

References

Alpaydin, E. 2014. Introduction to Machine Learning, 3rd Ed. Cambridge, MA: MIT Press, 640 pp.

Anderson, B.D., and J.B. Moore. 2012. Optimal Filtering. North Chelmsford, MA: Courier Corporation, 368 pp.

Anselin, L., and S.J. Rey. 2012. Spatial econometrics in an age of CyberGIScience. International Journal of Geographical Information Science 26(12):2211-2226.

Axtell, R., R. Axelrod, J.M. Epstein, and M.D. Cohen. 1996. Aligning simulation models: A case study and results. Computational and Mathematical Organization Theory 1(2):123-142.

Banerjee, S., B.P. Carlin, and A.E. Gelfand. 2014. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton, FL: CRC Press, 474 pp.

Barthélemy, M. 2011. Spatial networks. Physics Reports 499(1):1-101.

Bast, H., P. Brosi, and S. Storandt. 2014. Real-time movement visualization of public transit data. Pp. 331-340 in SIGSPATIAL'14, Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.

Bauer, P., A. Thorpe, and G. Brunet. 2015. The quiet revolution of numerical weather prediction. Nature 525(7567):47-55.

Bayarri, M.J., J.O. Berger, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R.J. Parthasarathy, R. Paulo, J. Sacks, and D. Walsh. 2007. Computer model validation with functional output. Annals of Statistics 35(5):1874-1906.

Beaumont, M.A. 2010. Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 41:379-406.

Berkooz, G., P. Holmes, and J.L. Lumley. 1993. The proper orthogonal decomposition in the analysis of turbulent flows. Annual Review of Fluid Dynamics 25:539-575.

Benner, P., S. Gugercin, and K. Willcox. 2015. A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Review 57(4):483-531.

Bishop, C.M. 2006. Pattern Recognition and Machine Learning. New York: Springer, 738 pp.

Bonabeau, E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences of the United States of America 99(Suppl 3):7280-7287.

Borgatti, S., K.M. Carley, and D. Krackhardt. 2006. Robustness of centrality measures under conditions of imperfect data. Social Networks 28(2):124-136.

Box, G.E.P., and D.W. Behnken. 1960. Some new three level designs for the study of quantitative variables. Technometrics 2(4):455-475.

Box, G.E.P., and K.B. Wilson. 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society Series B 13(1):1-45.

Box, G.E., G.M. Jenkins, G.C. Reinsel, and G.M. Ljung. 2015. Time Series Analysis: Forecasting and Control. New York: John Wiley & Sons, 712 pp.

Brandes, U., M. Eiglsperger, J. Lerner, and C. Pich. 2013. Graph Markup Language (GraphML). Pp. 517-541 in Handbook of Graph Drawing and Visualization, edited by R. Tamassia. Boca Raton, FL: CRC Press [online]. Available at https://cs.brown.edu/~rt/gdhandbook/chapters/graphml.pdf [accessed May 4, 2016].

Brooks, S., A. Gelman, G. Jones, and X.L. Meng, eds. 2011. Handbook of Markov Chain Monte Carlo. Boca Raton, FL: CRC Press.

Burton, R.M., and B. Obel. 1995. The validity of computational models in organization science: From model realism to purpose of the model. Computational and Mathematical Organization Theory 1(1):57-71.

Butler, M.P., P.M. Reed, K. Fisher-Vanden, K. Keller, and T. Wagener. 2014. Identifying parametric controls and dependencies in integrated assessment models using global sensitivity analysis. Environmental Modelling & Software 59:10-29.

Butler, T., and D. Estep. 2013. A numerical method for solving a stochastic inverse problem for parameters. Annals of Nuclear Energy 52:86-94.

Caminade, C., S. Kovats, J. Rocklov, A.M. Tompkins, A.P. Morse, F. Jesús Colón-González, H. Stenlund, P. Martens, and S.J. Lloyd. 2014. Impact of climate change on global malaria distribution. Proceedings of the National Academy of Sciences of the United States of America 111(9):3286-3291.

Cao, G., S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, and K. Soltani. 2015. A scalable framework for spatiotemporal analysis of location-based social media data. Computers, Environment and Urban Systems 51:70-82.

Carley, K.M., M.K. Martin, and B. Hirshman. 2009. The etiology of social change. Topics in Cognitive Science 1(4):621-650.

Carley, K.M., M.W. Bigrigg, and B. Diallo. 2012a. Data-to-model: A mixed initiative approach for rapid ethnographic assessment. Computational and Mathematical Organization Theory 18(3):300-327.

Carley, K.M., G. Morgan, M. Lanham, and J. Pfeffer. 2012b. Multi-modeling and sociocultural complexity. Pp. 128-137 in Advances in Design for Cross-Cultural Activities, Part II, edited by D.D. Schmorrow, and D.M. Nicholson. Boca Raton, FL: CRC Press.

Carley, K.M., J. Pfeffer, F. Morstatter, and H. Liu. 2014. Embassies burning: Toward a near-real -time assessment of social media using geo-temporal dynamic network analytics. Social Network Analysis and Mining 4:195.

Catlett, C., W.E. Allcock, P. Andrews, and 91 others. 2007. TeraGrid: Analysis of organization, system architecture, and middleware enabling new types of applications. Pp. 225-249 in High Performance Computing and Grids in Action, edited by Lucio Grandinetti. Advances in Parallel Computing Series, Vol. 16. Amsterdam: IOS Press.

Chancellor, E. 1999. Devil Take the Hindmost: A History of Financial Speculation. New York: Plume.

Chesshire, J.H., and A.J. Surrey. 1975. World energy resources and the limitations of computer modelling. Long Range Planning 8(3):54-61.

Chien, A.A., and V. Karamcheti. 2013. Moore's Law: The first ending and a new beginning. Computer 12(46):48-53.

Chinesta, F., A. Huerta, G. Rozza, and K. Willcox. 2016. Model reduction methods. Encyclopedia of Computational Mechanics, Vol. 2: Solids and Structures, 2nd Ed, Edited by E. Stein, R. de Borst, and Thomas T.J.R. Hughes. New York: John Wiley & Sons.

Clapper, J.R. 2015. Worldwide Threat Assessment of the U.S. Intelligence Community. Statement for the Record, Senate Select Committee on Intelligence, February 26, 2015.

Collins, W.D., A.P. Craig, J.E. Truesdale, A.V. Di Vittorio, A.D. Jones, B. Bond-Lamberty, K.V. Calvin, J.A. Edmonds, S.H. Kim, A.M. Thomson, P. Patel, Y. Zhou, J. Mao, X. Shi, P.E. Thornton, L.P. Chini, and G.C. Hurtt. 2015. The integrated Earth system model version 1: Formulation and functionality. Geoscientific Model Development 8(7):2203-2219.

Conti, S., J.P. Gosling, J.E. Oakley, and A. O'Hagan. 2009. Gaussian process emulation of dynamic computer codes. Biometrika 96(3):663-676.

Courtier, P., J.N. Thépaut, and A. Hollingsworth. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. Quarterly Journal of the Royal Meteorological Society 120(519):1367-1387.

Cressie, N., and C.K. Wikle. 2015. Statistics for Spatio-Temporal Data. New York: John Wiley & Sons, 512 pp.

Dahan-Dalmedico, A. 2001. History and epistemology of models: Meteorology (1946–1963) as a case study. Archive for History of Exact Sciences 55:395-422.

Das Sarma, A., H. Lee, H. Gonzalez, J. Madhavan, and A. Halevy. 2012. Efficient spatial sampling of large geographical tables. Pp. 193-204 in SIGMOD '12, Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data.

Davis, G.B., J. Olson, and K.M. Carley. 2008. OraGIS and Loom: Spatial and Temporal Extensions to the ORA Analysis Platform. Carnegie Mellon University, Institute for Software Research, Technical Report, CMU-ISR-08-121. Reprinted as DTIC ADA486288, June 2008.

Davis, P.K., and R.H. Anderson. 2004. Improving the composability of DoD models and simulations. Journal of Defense Modeling and Simulation: Applications, Methodology, Technology 1(1):5-17.

Davis, P.K., and J.H. Bigelow. 2003. Motivated Metamodels: Synthesis of Cause-Effect Reasoning and Statistical Metamodeling. RAND/MR-1570. Santa Monica, CA: RAND Corporation [online]. Available at http://www.rand.org/pubs/monograph_reports/MR1570.html [accessed September 26, 2016].

Davis, P.K., R.D. Shaver, and J. Beck. 2008. Portfolio-Analysis Methods For Assessing Capability Options. RAND Corporation [online]. Available at http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG662.pdf [accessed September 26, 2016].

Dean, J., and S. Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. Communications of the ACM 51(1):107-113.

Delli Gatti, D., S. Desiderio, E. Gaffeo, P. Cirillo, and M. Gallegati. 2011. Macroeconomics from the Bottom Up. Milan: Springer.

DOE (U.S. Department of Energy). 2009. Science Challenges and Future Directions: Climate Change Integrated Assessment Research. Report PNNL-18417 [online]. Available at http://science.energy.gov/~/media/ber/pdf/ia_workshop_low_res_06_25_09.pdf [accessed May 5, 2016].

Drouet, L., V. Bosetti, and M. Tavoni. 2015. Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. Nature Climate Change 5:937-940.

Edwards, P. 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, MA: MIT Press.

Efron, B., and C. Morris. 1977. Stein's paradox in statistics. Scientific American 236:119-127.

Egan, J.R., I.M. Hall, D.J. Lemon, and S. Leach. 2011. Modeling Legionnaires' disease outbreaks: Estimating the timing of an aerosolized release using symptom-onset dates. Epidemiology 22(2):188-198.

Eldawy, A., and M.F. Mokbel. 2015. The era of big spatial data: A survey. Information and Media Technologies 10(2):305-316.

Eldred, M.S., B.M. Adams, D.M. Gay, L.P. Swiler, K. Haskell, W.J. Bohnhoff, J.P. Eddy, W.E. Hart, J.P. Watson, J.D. Griffin, P.D. Hough, T.G. Kolda, P.J. Williams, and M.L. Martinez-Canales. 2007. DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 4.1 Reference Manual. SAND2006-4055. Albuquerque, NM: Sandia National Laboratories [online]. Available at https://dakota.sandia.gov/content/sand-reports [accessed May 5, 2016].

Evensen, G. 2009. Data Assimilation: The Ensemble Kalman Filter. Berlin: Springer.

FAO (Food and Agriculture Organization). 2012. The State of Food Insecurity in the World: Economic Growth is Necessary but Not Sufficient to Accelerate Reduction of Hunger and Malnutrition. Rome: FAO [online]. Available at http://www.fao.org/docrep/016/i3027e/i3027e.pdf [accessed May 5, 2016].

Fearnhead, P., and D. Prangle. 2012. Constructing summary statistics for approximate Bayesian computation: Semi-

automatic approximate Bayesian computation. Journal of the Royal Statistical Society Series B: Statistical Methodology 74(3):419-474.

Ferguson, N.M., M.J. Keeling, W.J. Edmunds, R. Gani, B.T. Grenfell, R.M. Anderson, and S. Leach. 2003. Planning for smallpox outbreaks. Nature 425(6959):681-685.

Fischoff, B., and A.L. Davis. 2014. Communicating scientific uncertainty. Proceedings of the National Academy of Sciences of the United States of America 111:13,664-13,671.

Forrester, J.W. 1961. Industrial Dynamics. Waltham, MA: Pegasus Communications, 464 pp.

Frantz, T.L., M. Cataldo, and K.M. Carley. 2009. Robustness of centrality measures under uncertainty: Examining the role of network topology. Computational and Mathematical Organization Theory 15(4):303-328.

Fryxell, B., K. Olson, P. Ricker, F.X. Timmes, M. Zingale, D.Q. Lamb, P. MacNeice, R. Rosner, J.W. Truran, and H. Tufo. 2000. FLASH: An adaptive mesh hydrodynamics code for modeling astrophysical thermonuclear flashes. Astrophysical Journal Supplement Series 131(1):273-334.

Gallup, J.L., and J.D. Sachs. 2001. The economic burden of malaria. American Journal of Tropical Medicine and Hygiene 64(1-2 Suppl):85-96.

Geanakoplos, J., R. Axtell, D.J. Farmer, P. Howitt, B. Conlee, J. Goldstein, M. Hendrey, N.M. Palmer, and C.-Y. Yang. 2012. Getting at systemic risk via an agent-based model of the housing market. American Economic Review 102(3):53-58.

Gelman, A., and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, UK: Cambridge University Press.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. Bayesian Data Analysis, 3rd Ed. Boca Raton, FL: Chapman and Hall/CRC Press, 675 pp.

Gething, P.W., D.L. Smith, A.P. Patil, A.J. Tatem, R.W. Snow, and S.I. Hay. 2010. Climate change and the global malaria recession. Nature 465(7296):342-345.

Gilbert, N. 2008. Agent-Based Models. Series: Quantitative Applications in the Social Sciences. Los Angeles, CA: SAGE Publications, 112 pp.

Giorgi, F., and L.O. Mearns. 1991. Approaches to the simulation of regional climate change: A review. Reviews of Geophysics 29:191-216.

Gutmann, E.G., T. Pruitt, M.P. Clark, L. Brekke, J.R. Arnold, D.A. Raff, and R.M. Rasmussen. 2014. An intercomparison of statistical downscaling methods used for water resource assessments in the United States. Water Resources Research 50:7167-7186.

Hanasaki, N., S. Fujimori, T. Yamamoto, S. Yoshikawa, Y. Masaki, Y. Hijioka, M. Kainuma, Y. Kanamori, T. Masui, K. Takahashi, and S. Kanae. 2013. A global water scarcity assessment under shared socio-economic pathways—Part 2: Water availability and scarcity. Hydrological Earth System Science 17(7):2393-2413.

Hannigan, J., G. Hernandez, R.M. Medina, R. Roos, and P. Shakarian. 2013. Mining for spatially-near communities in geo-located social networks. Pp. 16-23 in Association for the Advancement of Artificial Intelligence—Social Networks and Social Contagion: Web Analytics and Computational Social Science, Arlington, VA, November 15-17, Technical Report.

Hansen, J., G. Jacobs, L. Hsu, J. Dykes, J. Dastugue, R. Allard, C. Barron, D. Lalejini, M. Abramson, S. Russell, and R. Mittu. 2011. Information domination: Dynamically coupling METOC and INTEL for improved guidance for piracy interdiction. NRL Review (2011):109-119.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed. New York: Springer-Verlag, 763 pp.

Hay, S.I., C.A. Guerra, P.W. Gething, A.P. Patil, A.J. Tatem, A.M. Noor, C.W. Kabaria, B.H. Manh, I.R.F. Elyazar, S. Brooker, D.L. Smith, R.A. Moyeed, and R.W. Snow. 2009. A world malaria map: Plasmodium falciparum endemicity in 2007. PLoS Medicine 6(3):286-301.

Hejazi, M.I., J. Edmonds, L. Clarke, P. Kyle, E. Davies, V. Chaturvedi, M. Wise, P. Patel, J. Eom, and K. Calvin.

2014. Integrated assessment of global water scarcity over the 21st century under multiple climate change mitigation policies. Hydrology and Earth System Sciences 18(8):2859-2883.

Helton, J.C. 1993. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. Reliability Engineering & System Safety 42(2-3):327-367.

Hendricks, J.S., G.W. McKinney, M.L. Fensin, M.R. James, R.C. Johns, J.W. Durkee, J.P. Finch, D.B. Pelowitz, L.S. Waters, M.W. Johnson, and F.X. Gallmeier. 2008. MCNPX 2.6.0 Extensions. LA-UR-08-2216. Los Alamos National Laboratory [online]. Available at https://mcnpx.lanl.gov/opendocs/versions/v260/v260.pdf [accessed May 5, 2016].

Higdon, D. 2006. A primer on space-time modeling from a Bayesian perspective. In Statistical Methods for Spatio-Temporal Systems. Monographs on Statistics and Applied Probability. B. Finkenstadt, L. Held, and V. Isham, eds. Chapman and Hall, pp. 217-279.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley. 2008. Computer model calibration using high-dimensional output. Journal of the American Statistical Association 103(482):570-583.

Hofmann, M.A. 2004. Challenges of model interoperation in military simulations. Simulation 80(12):659-667.

Holland, J.H. 1992. Complex adaptive systems. Daedalus 121(1):17-33.

Holt, J. 2009. A summary of the primary causes of the housing bubble and the resulting credit crisis: A non-technical paper. Journal of Business Inquiry 8(1):120-129.

Howard, R.A. 1968. The foundations of decision analysis. IEEE Transactions on Systems Science and Cybernetics 4(3):211-219.

Ide, K., P. Courtier, M. Ghil, and A.C. Lorenc. 1997. Unified notation for data assimilation: Operational, sequential and variational. Journal of the Meteorological Society of Japan 75(1B):181-189.

IPCC (Intergovernmental Panel on Climate Change). 2013. Summary for policymakers. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley. Cambridge, UK: Cambridge University Press.

Jones, A.E., and A.P. Morse. 2012. Skill of ENSEMBLES seasonal re-forecasts for malaria prediction in West Africa. Geophysical Research Letters 39:L23707, DOI: 10.1029/2012gl054040.

Joseph, K., K.M. Carley, D. Filonuk, G.P. Morgan, and J. Pfeffer. 2014. Arab Spring: From news data to forecasting. Social Network Analysis and Mining 4(1):177, DOI: 10.1007/s13278-014-0177-5.

Kaipio, J., and E. Somersalo. 2006. Statistical and Computational Inverse Problems. Applied Mathematical Science Vol. 160. New York: Springer.

Kalman, R.E., and R.S. Bucy. 1961. New results in linear filtering and prediction theory. Journal of Basic Engineering 83(1):95-108.

Kas, M., K.M. Carley, and L.R. Carley. 2012. Who was where, when? Spatiotemporal analysis of researcher mobility in nuclear science. In Proceedings of the International Workshop on Spatio Temporal Data Integration and Retrieval (STIR 2012), April 1, 2012, Washington, DC.

Keeney, R.L. 1982. Decision analysis: An overview. Operations Research 30(5):803-838.

Kennedy, M.C., and A. O'Hagan. 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society Series B: Statistical Methodology 63(3):425-464.

Kiehl, J.T. 2007. Twentieth century climate model response and climate sensitivity. Geophysical Research Letters 34(22):L22710.

Kim, S.H., M. Hejazi, L. Liu, K. Calvin, L. Clarke, J. Edmonds, P. Kyle, P. Patel, M. Wise, and E. Davies. 2016. Balancing global water availability and use at basin scale in an integrated assessment model. Climatic Change 136(2):217-231.

Kirk, M. 2011. Ending Somali Piracy Against American and Allied Shipping [online]. Available at http://kirk.senate.gov/pdfs/KirkReportfinal2.pdf [accessed May 5, 2016].

Kogge, P., K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K.

Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R.S. Willians, and K. Yelick. 2008. Exascale Computing Study: Technology Challenges in Achieving Exascale Systems. DARPA, Information Processing Techniques Office, and Air Force Research Laboratory, 278 pp. [online]. Available at http://staff.kfupm.edu.sa/ics/ahkhan/Resources/Articles/ExaScale%20Computing/TR-2008-13.pdf [accessed May 5, 2016].

Kolaczyk, E.D., and G. Csárdi. 2014. Statistical Analysis of Network Data with R. Use R, Vol. 65. New York: Springer, 207 pp.

Kotamarthi, R., L. Mearns, K. Hayhoe, C.L. Castro, and D. Wuebble. 2016. Use of Climate Information for Decision-Making and Impacts Research: State of Our Understanding. Prepared for the Department of Defense, Strategic Environmental Research and Development Program. 55 pp.

Kou, Y., C.T., Lu, S. Sirwongwattana, and Y.P. Huang. 2004. Survey of fraud detection techniques. Pp. 749-754 in Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control, March 21-23, 2004, Taipei, Taiwan, Vol. 2. Piscataway, NJ: IEEE.

Krey, V. 2014. Global energy-climate scenarios and models: A review. Wiley Interdisciplinary Reviews Energy and Environment 3(4):363-383.

Lansing, J.S. 2003. Complex adaptive systems. Annual Review of Anthropology 32:183-204.

Lauderdale, J.M., C. Caminade, A.E. Heath, A.E. Jones, D.A. MacLeod, K.C. Gouda, U.S. Murty, P. Goswami, S.R. Mutheneni, and A.P. Morse. 2014. Towards seasonal forecasting of malaria in India. Malaria Journal 13:310.

Levis, A.H., K.M. Carley, and G. Karsai. 2011. Resilient Architectures for Integrated Command and Control in a Contested Cyber Environment. SAL/FR-11-02. Final Technical Report to the Air Force Research Laboratory, by George Mason University, Fairfax, VA [online]. Available at http://www.casos.cs.cmu.edu/publications/papers/RC2withSF298_Final2.pdf [accessed May 5, 2016].

Liu, J.S., and R. Chen. 1998. Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association 93(443):1032-1044.

Lorenz, E.N. 1963. On the predictability of hydrodynamic flow. Transactions of the New York Academy of Sciences 25(4):409-432.

Lorenz, J., H. Rauhut, F. Schweitzer, and D. Helbing. 2011. How social influence can undermine the wisdom of crowd effect. Proceedings of the National Academy of Sciences of the United States of America 108(22):9020-9025.

MacLeod, D.A., A. Jones, F. Di Giuseppe, C. Caminade, and A.P. Morse. 2015. Demonstration of successful malaria forecasts for Botswana using an operational seasonal climate model. Environmental Research Letters 10(4), DOI: 10.1088/1748-9326/10/4/044005.

Manabe, S., and K. Bryan. 1969. Climate calculations with a combined ocean-atmosphere model. Journal of the Atmospheric Sciences 26(4):786-789.

Manabe, S., and R.T. Wetherald. 1975. The effects of doubling the CO2 concentration on the climate of a general circulation model. Journal of Atmospheric Sciences 32(1):3-15.

Marchuk, G.I. 1995. Adjoint Equations and Analysis of Complex Systems. New York: Springer, 468 pp.

Martin, J., L.C. Wilcox, C. Burstedde, and O. Ghattas. 2012. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM Journal on Scientific Computing 34(3):A1460-A1487.

Marzouk, Y.M., and H.N. Najm. 2009. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. Journal of Computational Physics 228(6):1862-1902.

McGuire, T.R. 1997. The last northern cod. Journal of Political Ecology 4(1):41-54.

Medina, R.M., and G.F. Hepner. 2011. Advancing the understanding of sociospatial dependencies in terrorist networks. Transactions in GIS 15(5):577-597.

Medina, R.M., and G.F. Hepner. 2015. A note of the state of geography and geospatial intelligence research. NGA Pathfinder 13(1):8-9.

Mell, P., and T. Grance. 2011. The NIST Definition of Cloud Computing. NIST Special Publication 800-145. National Institute of Standards and Technology, 7 pp. [online]. Available at http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf [accessed May 5, 2016].

Merrill, J.A., B. Sheehan, K.M. Carley, and P.D. Stetson. 2015. Transition networks in a cohort of patients with congestive heart failure. A novel application of informatics methods to inform care coordination. Applied Clinical Informatics 6(3):548-564.

Mignolet, M.P., and C. Soize. 2008. Stochastic reduced order models for uncertain geometrically nonlinear dynamical systems. Computer Methods in Applied Mechanics and Engineering 197(45-48):3951-3963.

Miller, J.H. 1998. Active nonlinear tests (ANTs) of complex simulations models. Management Science 44(6):820-830.

Moore, L.M. 1981. Principle component analysis in linear systems: Controllability, observability, and model reduction. IEEE Transactions on Automatic Control 26(1):17-31.

Morgan, G.P., and K.M. Carley. 2012. Modeling formal and informal ties within an organization: A multiple model integration. Pp. 253-292 in The Garbage Can Model of Organizational Choice: Looking Forward at Forty, edited by A. Lomi and R. Harrison. Research in the Sociology of Organizations Vol. 36. Bingley, UK: Emerald Group Publishing.

Morgan, M.G., and M. Henrion. 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge, U.K.: Cambridge University Press, 332 pp.

Moss, R.H., J.A. Edmonds, K.A. Hibbard, M.R. Manning, S.K. Rose, D.P. van Vuuren, T.R. Carter, S. Emori, M. Kainuma, T. Kram, G.A. Meehl, G.F. Mitchell, N. Nakicenovic, K. Riahl, S.J. Smith, R.J. Stouffer, A.M. Thomson, J.P. Weyant, and T.J. Wilbanks. 2010. The next generation of scenarios for climate change research and assessment. Nature 463(7282):747-756.

Myers, M.F., D.J. Rogers, J. Cox, A. Flahault, and S.I. Hay. 2000. Forecasting disease risk for increased epidemic preparedness in public health. Advances in Parasitology 47:309-330.

NAC (NASA Advisory Council). 1986. Earth System Science: A Program for Global Change. Washington, DC: NASA.

NASEM (The National Academies of Sciences, Engineering, and Medicine). 2016a. Next Generation Earth System Prediction: Strategies for Sub-Seasonal to Seasonal Forecasts. Washington, DC: The National Academies Press.

NASEM. 2016b. Fostering Transformative Research in the Geographical Sciences. Washington, DC: The National Academies Press.

NIC (National Intelligence Council). 2012. Global Trends 2030: Alternative Worlds. NIC 2012-001 [online]. Available at https://cgsr.llnl.gov/content/assets/docs/Global_Trends_2030-NIC-US-Dec12.pdf [accessed May 9, 2016].

North, M.J., and C.M. Macal. 2007. Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford: Oxford University Press.

NRC (National Research Council). 1979. Carbon Dioxide and Climate: A Scientific Assessment. Washington, DC: National Academy Press, 22 pp.

NRC. 2006. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts. Washington, DC: The National Academies Press, 124 pp.

NRC. 2007a. Review of the U.S. Climate Change Science Program's Synthesis and Assessment Product 5.2, "Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Climate Decision Making." Washington, DC: The National Academies Press, 64 pp.

NRC. 2007b. Models in Environmental Regulatory Decision Making. Washington, DC: The National Academies Press, 286 pp.

NRC. 2012. Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification. Washington, DC: The National Academies Press, 184 pp.

NRC. 2013. Future U.S. Workforce for Geospatial Intelligence. Washington, DC: The National Academies Press, 172 pp.

Oakley, J., and A. O'Hagan. 2002. Bayesian inference for the uncertainty distribution of computer model outputs. Biometrika 89(4):769-784.

Oberkampf, W.L., and C.J. Roy. 2010. Verification and Validation in Scientific Computing. Cambridge, UK: Cambridge University Press, 790 pp.

Oberkampf, W.L., T.G. Trucano, and C. Hirsch. 2004. Verification, validation, and predictive capability in computational engineering and physics. Applied Mechanical Reviews 57(5):345-384.

O'Hagan, A. 2006. Bayesian analysis of computer code outputs: A tutorial. Reliability Engineering & System Safety 91(10-11):1290-1300.

Oliver, D., S. Shekhar, J.M. Kang, R. Laubscher, V. Carlan, and A. Bannur. 2014. A K-Main Routes approach to spatial network activity summarization. IEEE Transactions on Knowledge and Data Engineering 26(6):1464-1478.

Oliver, T.A., G. Terejanu, C.S. Simmons, and R.D. Moser. 2015. Validating predictions of unobserved quantities. Computer Methods in Applied Mechanics and Engineering 283(1):1310-1335.

Olson, J.F., and K.M. Carley. 2008. Summarization and information loss in network analysis. In Workshop on Link Analysis, Counter-terrorism, and Security, held in conjunction with the SIAM International Conference on Data Mining (SDM), April 2008.

Olson, J., and K.M. Carley. 2009. Visualizing Spatial Dependencies in Network Topology. Carnegie Mellon University, Institute for Software Research, Technical Report CMU-ISR-09-127. Reprinted as DTIC ADA525370, July 12, 2010.

Oort, A.H., and E.M. Rasmusson. 1970. On the annual variation of the monthly mean meridional circulation. Monthly Weather Review 98:423-442.

Ott, E., B.R. Hunt, I. Szunyogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil, and J.A. Yorke. 2004. A local ensemble Kalman filter for atmospheric data assimilation. Tellus A 56(5):415-428.

Pace, D.K. 2004. Modeling and simulation verification and validation challenges. Johns Hopkins APL Technical Digest 25(2):163-172.

Pascual, M., and M.J. Bouma. 2009. Do rising temperatures matter? Ecology 90(4):906-912.

Pavlis, N.K., S.A. Holmes, S.C. Kenyon, and J.K. Factor. 2012. The development and evaluation of the Earth Gravitational Model 2008 (EGM2008). Journal of Geophysical Research 117: B04406, DOI:10.1029/2011JB008916.

Phillips, N.A. 1956. The general circulation of the atmosphere: A numerical experiment. Quarterly Journal of the Royal Meteorological Society 82(352):123-164.

Platzman, G.W. 1979. The ENIAC computations of 1950: Gateway to numerical weather prediction. Bulletin of the American Meteorological Society 60(4):302-312.

Raftery, A.E., T. Gneiting, F. Balabdaoui, and M. Polakowski. 2005. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review 133(5):1155-1174.

Raiffa, H. 1968. Decision Analysis: Introductory Lectures on Choices under Uncertainty. Reading, MA: Addison-Wesley.

Reed, D.A., and J. Dongarra. 2015. Exascale computing and big data. Communications of the ACM 58(7):56-68.

Reichler, T., and J. Kim. 2008. How well do coupled models simulate today's climate? Bulletin of the American Meteorological Society 89(3):303-312.

Richardson, L.F. 1922. Weather Prediction by Numerical Process. Cambridge, UK: Cambridge University Press.

Ristic, B., S. Arulampalam, and N.J. Gordon. 2004. Beyond the Kalman Filter: Particle Filters for Tracking Applications. Boston, MA: Artech House.

Roache, P.J. 2002. Code verification by the method of manufactured solutions. Journal of Fluids Engineering 124(1):4-10.

Rosoff, H., and D. von Winterfeldt. 2007. A risk and economic analysis of dirty bomb attacks on the ports of Los Angeles and Long Beach. Risk Analysis 27(3):533-546.

Rutter, M. 2007. Proceeding from observed correlation to causal inference: The use of natural experiments. Perspectives on Psychological Science 2(4):377-395.

Ruttimann, J. 2006. 2020 computing: Milestones in scientific computing. Nature 440(7083):399-405.

Sacks, J., W.J. Welch, T.J. Mitchell, and H.P. Wynn. 1989. Design and analysis of computer experiments. Statistical Science 4(4):409-423.

Sadilek, A., H.A. Kautz, and V. Silenzio. 2012. Predicting disease transmission from geo-tagged micro-blog data. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 136-142.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. Global Sensitivity Analysis: The Primer. Chichester, UK: John Wiley & Sons, 304 pp.

Schapire, R.E. 2003. The boosting approach to machine learning: An overview. Pp. 149-171 in Nonlinear Estimation and Classification, edited by D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick, and B. Yu. Lecture Notes in Statistics Vol. 171. New York: Springer.

Schlosser, C.A., K.M. Strzepek, X. Gao, A. Gueneau, C. Fant, S. Paltsev, B. Rasheed, T. Smith-Greico, É. Blanc, H.D. Jacoby, and J.M. Reilly. 2014. The future of global water stress: An integrated assessment. Earths Future 2(8):341-361.

Scott, J. 2013. Social Network Analysis, 3rd Ed. Los Angeles, CA: SAGE, 216 pp.

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1):1-47.

Shekhar, S., S. Ravada, V. Kumar, D. Chubb, and G. Turner. 1996. Parallelizing a GIS on a shared address space architecture. IEEE Computer 29(12):42-48.

Shekhar, S., D. Chubb, and G. Turner. 1998. Declustering and load-balancing methods for parallelizing geographic information systems. IEEE Transactions on Knowledge and Data Engineering 10(4):632-655.

Shekhar, S., M.R. Evans, J.M. Kang, and P. Mohan. 2011. Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3):193-214.

Shekhar, S., Z. Jiang, R.Y. Ali, E. Eftelioglu, X. Tang, V.M.V. Gunturi, and X. Zhou. 2015. Spatiotemporal data mining: A computational perspective. ISPRS International Journal on Geo-Information 4(4):2306-2338.

Shivaji, T., C. Sousa Pinto, A. San-Bento, L.A. Oliveira Serra, J. Valente, J. Machado, T. Marques, L. Carvalho, P.J. Nogueira, B. Nunes, and P. Vasconcelos. 2014. A large community outbreak of Legionnaires' disease in Vila Franca de Xira, Portugal, October to November 2014. Eurosurveillance 19(50):Article 3 [online]. Available at http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20991 [accessed May 9, 2016].

Simmons, A.J., and A. Hollingsworth. 2002. Some aspects of the improvement in skill of numerical weather prediction. Quarterly Journal of the Royal Meteorological Society Part B 128(580):647-677.

Siraj, A.S., M. Santos-Vega, M.J. Bouma, D. Yadeta, D. Ruiz Carrascal, and M. Pascual. 2014. Altitudinal changes in malaria incidence in highlands of Ethiopia and Colombia. Science 343(6175):1154-1158.

Slingo, J., and T. Palmer. 2011. Uncertainty in weather and climate prediction. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 369(1956):4751-4767.

Slootmaker, L.A., E. Regnier, J.A. Hansen, and T.W. Lucas. 2013. User focus and simulation improve predictions of piracy risk. Interfaces 43(3):256-267.

Smith, R.C. 2013. Uncertainty Quantification: Theory, Implementation, and Applications. Philadelphia: Society for Industrial and Applied Mathematics, 383 pp.

Smith, R.L., C. Tebaldi, D. Nychka, and L.O. Mearns. 2009. Bayesian modeling of uncertainty in ensembles of climate models. Journal of the American Statistical Association 104(485):97-116.

Snow, J. 1855. On the Mode of Communication of Cholera. London: John Churchill [online]. Available at http://www.ph.ucla.edu/epi/snow/snowbook.html [accessed May 9, 2016].

Spanos, P.D., and R. Ghanem. 1989. Stochastic finite element expansion for random media. Journal of Engineering Mechanics 115(5):1035-1053.

Sterman, J.D. 2000. Business Dynamics: Systems Thinking and Modeling for a Complex World. New York: McGraw-Hill, 982 pp.

Stigler, S.M. 1986. The English breakthrough: Galton. Pp. 265-299 in The History of Statistics: The Measurement of Uncertainty before 1900. Cambridge, MA: Harvard University Press.

Stroeve, J.C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W.N. Meier. 2012. Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. Geophysical Research Letters 39, L16502, DOI:10.1029/2012GL052676.

Sung, K.K., and T. Poggio. 1998. Example-based learning for view-based human face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(1):39-51.

Tang, L.-A., Y. Zheng, J. Yuan, J. Han, A. Leung, C.-C. Hung, and W.-C. Peng. 2012. On discovery of traveling companions from streaming trajectories. In ICDE 2012, Proceedings of the IEEE 28th International Conference on Data Engineering. April 1-5, 2012, Washington, DC.

Tang, W., and S. Wang. 2009. HPABM: A hierarchical parallel simulation framework for spatially-explicit agent-based models. Transactions in GIS 13(3):315-333.

Tarantola, A. 2005. Inverse Problem Theory and Methods for Model Parameter Estimation. Philadelphia, PA: Society for Industrial and Applied Mathematics [online]. Available at http://www.ipgp.fr/~tarantola/Files/Professional/Books/InverseProblemTheory.pdf [accessed May 9, 2016].

Taylor, J.B. 2009. The Financial Crisis and the Policy Responses: An Empirical Analysis of What Went Wrong. Working Paper No. 14631. Cambridge, MA: National Bureau of Economic Research.

Tebaldi, C., and R. Knutti. 2007. The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 365(1857):2053-2075.

Thomson, M.C., F.J. Doblas-Reyes, S.J. Mason, R. Hagedorn, S.J. Connor, T. Phindela, A.P. Morse, and T.N. Palmer. 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439(7076):576-579.

Tippett, M.K., J.L. Anderson, C.H. Bishop, T.M. Hamill, and J.S. Whitaker. 2003. Ensemble square root filters. Monthly Weather Review 131(7):1485-1490.

Tolk, A., and J.A. Muguira. 2003. The levels of conceptual interoperability model. In Proceedings of the 2003 Fall Simulation Interoperability Workshop 7, pp. 1-11.

Tompkins, A.M., and F. Di Giuseppe. 2015. Potential predictability of malaria in Africa using ECMWF monthly and seasonal climate forecasts. Journal of Applied Meteorology and Climatology 54(3):521-540.

Tsvetovat, M., J. Reminga, and K.M. Carley. 2003. DyNetML: Interchange format for rich social network data. Proceedings of the NAACSOS (North American Association for Computational Social and Organizational Sciences) Conference 2003, June 22-25, Pittsburgh, PA [online]. Available at http://www.casos.cs.cmu.edu/events/conferences/2003/proceedings.html. http://www.casos.cs.cmu.edu/publications/papers/tsvetovat_2003_dynetmlinterchange.pdf [accessed May 10, 2016].

Van Holt, T., J.C. Johnson, J. Brinkley, K.M. Carley, and J. Caspersen. 2012. Structure of ethnic violence in Sudan: An automated content, meta-network and geospatial analytical approach. Computational and Mathematical Organization Theory 18:340-355.

Wang, S., and M.P. Armstrong. 2009. A theoretical approach to the use of cyberinfrastructure in geographical analysis. International Journal of Geographical Information Science 23(2):169-193.

Wang, S., and X.-G. Zhu. 2008. Coupling cyberinfrastructure and Geographic Information Systems to empower ecological and environmental research. BioScience 58(2):94-95.

Wang, S., H. Hu, T. Lin, Y. Liu, A. Padmanabhan, and K. Soltani. 2014. CyberGIS for data-intensive knowledge discovery. ACM SIGSPATIAL Newsletter 6(2):26-33.

Wang, S., Y. Liu, and A. Padmanabhan. In press. Open cyberGIS software for geospatial research and education in the big data era. SoftwareX, DOI:10.1016/j.softx.2015.10.003.

Washington, W.M., L. Buja, and A. Craig. 2009. The computational future for climate and Earth system models: On the path to petaflop and beyond. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367(1890):833-846.

Wasserman, S., and K. Faust. 1994. Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences, Vol. 8. Cambridge, UK: Cambridge University Press.

Weinberger, S. 2011. Web of war. Nature 471:566-568.

West, M., and J. Harrison. 1999. Bayesian Forecasting and Dynamic Models, 2nd Ed. New York: Springer, 682 pp.

WHO (World Health Organization). 2012. World Malaria Report 2012. Geneva: WHO, 249 pp.

Wikle, C.K., and M.B. Hooten. 2010. A general science-based framework for dynamical spatio-temporal models. Test 19(3):417-451.

Wilby, R.L., S.P. Charles, E. Zorita, B. Timbal, P. Whetton, and L.O. Mearns. 2004. Guidelines for the use of climate scenarios developed from statistical downscaling methods. Available at http://ipcc-ddc.cru.uea.ac.uk/guidelines/dgm_no2_v1_09_2004.pdf [accessed September 26, 2016].

Wilhelmi, O., J. Boehnert, and K. Sampson. 2016. Visualizing the climate's future. Eos 97, DOI:10.1029/2016EO042207.

Willcox, K., and J. Peraire. 2002. Balanced model reduction via the proper orthogonal decomposition. AIAA Journal 40(11):2323-2330.

Wright, C., and T. Cheetham. 1999. The strengths and limitations of parental heights as a predictor of attained height. Archives of Disease in Childhood 81(3):257-260.

Yelick, K., S. Coghlan, B. Draney, and R.S. Canon. 2011. The Magellan Report on Cloud Computing for Science. U.S. Department of Energy [online]. Available at http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf [accessed May 10, 2016].

Zeigler, B.P., H. Praehofer, and T.G. Kim. 2000. Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems, 2nd Ed. New York: Academic Press, 510 pp.

Zhong, C., S.M. Arisona, X. Huang, M. Batty, and G. Schmitt. 2014. Detecting the dynamics of urban structure through spatial network analysis. International Journal of Geographical Information Science 28(11):2178-2199.

Zhou, X., S. Shekhar, and R.Y. Ali. 2014. Spatio-temporal change footprint pattern discovery: An interdisciplinary survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4:1-23.

Appendix A

Combining Models

Multiple, diverse models of components or subsystems are commonly linked together to form a collective model of a complex system. The use of multiple models, often referred to as multimodeling, reduces the time to develop models, supports validation, and enables multiple teams to take part in the exercise (Carley et al., 2012b). On the other hand, their use requires some special considerations. For example, multimodeling can be facilitated by the use of simulation testbeds, workflow systems, and infrastructure that supports linking models together. However, existing tools presume a common level of temporal and spatial resolution for all models, which is generally not the case, and the tools do not consider group resolution level, which is important in social system and geonetwork models. Hence, existing testbeds are insufficient for linking many types of models. In addition, the use of a model outside of the investigation for which it was designed can lead to serious misinterpretations of the results. This misinterpretation can be worse in a multimodeling framework, because most models will be used outside of their original contexts.

This appendix describes five methods for linking models (docking, collaboration, interoperation, integration, and coupling), and requirements for developing and using combined models of complex systems.

METHODS FOR LINKING MODELS

Analysts often need to rapidly analyze complex systems and answer diverse and changing questions about those systems. In principle, this functionality could be provided either in a unified model and analysis package, or by a federation of models and tools. For a federated system, all models and tools need to be made interoperable through the use of standardized data formats and exchange languages, and the alignment of the temporal, spatial, and grouping processes. The set of approaches commonly used to meet these needs—docking, collaboration, interoperation, integration, and coupling—are described below. The different approaches are not mutually exclusive, and two or more may be used in a model-based investigation or in composing a new system out of component models.

Docking

For computational models, docking refers to a type of alignment in which the docked models share some of the same input and generate some of the same output (Axtell et al., 1996; Burton and Obel, 1995; see Figure A.1). If two or more models can be shown to generate comparable output from the same input, they are docked. When two or more models are docked, they share some input but may each have additional unique input. In this case, all models that are docked will generate some output that is comparable and possibly some that is unique to the model. When inputs and outputs are the same, the internal processes in the model are aligned and the models are said to be docked. Within this scope, any differences in the theories inherent in the models or the details of the way they are programmed do not matter.

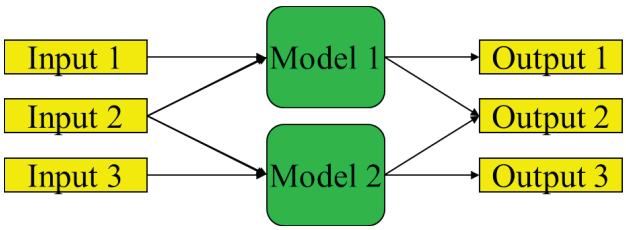


FIGURE A.1 Docking. SOURCE: Carley et al. (2012b).

Advantages of docking are that it is economical, flexible, and extensible to many models. For sociotechnical models, where the traditional methods of validation do not apply, docking is viewed as a form of validation. The idea is that any one theory or instantiation of a theory might be wrong, but if multiple models with diverse theoretical perspectives are docked, a robust and possibly overdetermined relation exists and is being captured by the suite of models. Finally, docking supports aligning one model with another (Axtell et al., 1996; Burton and Obel, 1995), which clarifies what the modeling assumptions are and how they affect the results. Model alignment helps the user understand the tradeoffs and risks in using one model over another and brings to light the limitations in each of the docked models. Disadvantages of docking are that the procedure requires participation by model developers or great documentation of the models, provides no theory development, and does not support reuse.

Collaboration

Collaboration is a federated approach in which two models sharing input data are used in conjunction to produce a collection of results that can be used together (see Figure A.2). The models must have some input in common, but each model may also use additional unique data. In this case, each model generates distinct outputs. Advantages of this approach are that it is economical, flexible, and extensible, is itself a form of validation (through triangulation), and supports the use of nonfused data. Disadvantages are that this approach requires moderate theory development and does not automatically guarantee reuse.

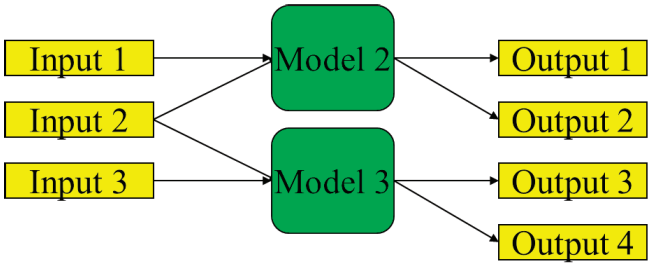


FIGURE A.2 Collaboration. SOURCE: Carley et al. (2012b).

Interoperation

Interoperation occurs when the outputs of one model are the inputs to another. In this case, the models form a chain that is implicitly the merger of theories (Levis et al., 2011; see Figure A.3). Interoperation has numerous advantages. Interoperation can occur at many levels (e.g., Tolk and Muguira, 2003). A federated system formed by interoperation is generally easy to use, extensible, flexible, highly reusable, and economical to develop, and it supports the development of theory. The key disadvantages are that making models interoperable often requires some code development and some theory development. For models composed out of parts, interoperation is easier if the models are in the same framework; at the same geographic, temporal, and group level of fidelity; and from the same theoretic tradition. In addition, the extent to which a composed model is valid depends on whether the component models are validated, whether the systems denoted by the two different models can be segregated, or whether complex interactions occur as the two systems interact. The science of model integration and the limits of composability are active areas of research (Davis and Anderson, 2004).

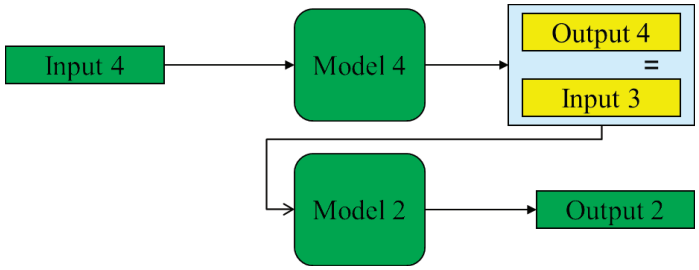


FIGURE A.3 Interoperation. SOURCE: Modified from Carley et al. (2012b).

Integration

Integration produces a single monolithic system in which all submodels are incorporated into a single model, often by a single programmer (Morgan and Carley, 2012). In such a system, all inputs go into the monolithic model, which, in turn, generates all outputs (see Figure A.4). Monolithic packages can provide great analytic power. However, they are often highly complex and require specialized functionalities or metrics, which can create a steep learning curve for analysts.

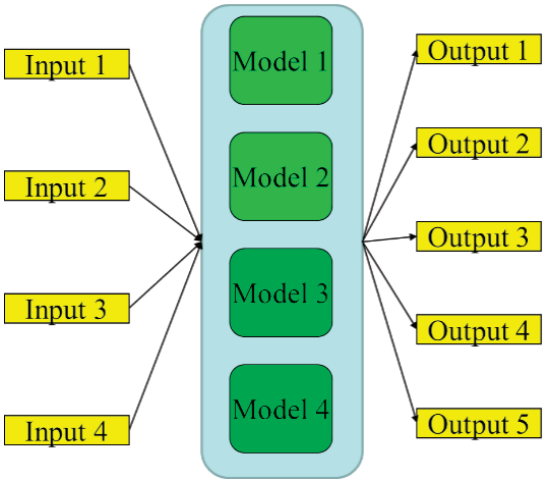


FIGURE A.4 Integration. SOURCE: Carley et al. (2012b).

An advantage of integration is that the overall system may be easier to use and maintain because there is a single user interface and updates and changes are controlled by a single group. The model might also run faster because there are no duplicate subparts, and the overall system can be optimized. Finally, developing a monolithic system can lead to new theory development. The disadvantages are that monolithic systems are costly to develop and extend, difficult to validate and are never completely validated, have low extensibility and low flexibility, require extensive new code development, and are unlikely to be reused.

Integrated models can be modular. However, the difference between a modular integrated system and a system formed through interoperation and docking has to do, in part, with the pedigree of the models and the ease of bringing in new submodels. In a composable system, each of the modules may be from a different theoretical tradition, built by different model teams, and separately validated. In an integrated system, the modules are typically designed from a software performance perspective and are not driven by diverse theoretical conceptions. Integration is a more top-down approach to developing a complex model, whereas interoperation via docking and the other methods mentioned above is a more bottom-up approach. Integrated systems are typically controlled centrally and it is often more difficult to add new modules from radically different perspectives than it is to add them in bottom-up systems. In principle, both the bottom-up and top-down approaches could result in the same design. However, for simulations of sociotechnical systems, theoretical understanding is still relatively weak, and so the two approaches tend to produce different results. The integrative approach typically results in an early hardening of the model design, which prevents the integration of radical departures from early conceptualizations as new discoveries are made.

Coupling

In some situations, linking the models is insufficient because the models depend on one other. Combining dependent models is referred to as coupling. For example, heat is transported from the tropics to the poles by atmospheric circulation and ocean currents. These motions are not independent, but they are strongly coupled as the atmosphere and ocean exchange heat, water, momentum, and biochemical species. Coupled modeld are typically used to understand interactions between subsystems.

The development of coupled models requires an understanding of all components of the system being coupled and all the modeling paradigms being used (Zeigler et al., 2000). Whereas disciplinary experts may engage more heavily in the development of a subsystem model, those charged with integrating subsystems must understand the

various subsystems and the way they are represented in the models. Typically, developing subsystem models with the intent of coupling ensures that common variables, units, spatial resolution, and temporal resolution are used and that the outputs produced by one subsystem are directly usable by another subsystem. Subsystem models that are developed independently and later coupled often require the development of methods for translating information from one subsystem to another or a lengthy redevelopment of one or more components. Developers often retain the ability to operate each subsystem individually, a configuration that allows faster run time when the question of interest is specific to an individual subsystem.

Coupled models are often developed with an interface that handles the transfer of information between subsystem models. For example, the Community Earth System Model[1] has a "coupler" that passes information from one component to another (e.g., precipitation is passed from the atmosphere component to the land component). The Global Change Assessment Model[2] uses a marketplace to pass information from one component to another (e.g., supply of bioenergy is passed from the agriculture module to the energy module via the marketplace).

There are many challenges in developing and using coupled models. First, these models require interactions between different communities and models, which often have different languages and methods of operating. Second, coupled models are often computationally expensive. As a result, developers are often faced with a choice between including a complex subsystem model, developing an emulator of that subsystem, or excluding the subsystem. Third, coupled models include numerous processes and feedbacks, and so analysis of results can be challenging. Experiments need to be structured carefully to isolate the effect of different mechanisms on various systems.

Choosing an Approach to Linking Models

How models are linked together depends in part on their alignment. Three key dimensions need to be aligned: spatial, temporal, and organizational/group. At the spatial level, the issue is whether the models cover the same spatial region, at the same level of granularity. For example, two city-level models of Morocco are highly aligned, whereas one model at the block level and one at the state level, or one generic and one of Saudi Arabia, are highly unaligned. At the temporal level, the issue is whether the models cover the same time span, at the same level of granularity. For example, two monthly models for 2000 are highly aligned, whereas one model at the day level in 1990 and one at the year level from 2000 to 2015 are highly unaligned. At the organizational level, the issue is whether the models cover the collectives of individuals, at the same level of granularity. For example, two models of individuals as separate agents are highly aligned, whereas one model at the small group level and one at the organizational level are highly unaligned.

Geographic, temporal, and organizational alignment are necessary but may not be sufficient. For example, models might not be in alignment because of differences in the functions or processes used. One agent-based model might maximize a trivial utility function, whereas another might look at more complex matters and reflect cognitive biases. Even if aligned in space, time, and group, they would not be aligned cognitively.

The following questions can be used to help identify the best approach for combining models:

- *Are the models operating in the same time window?*
  - Yes: docking, interoperation, or integration
  - No: collaboration or interoperation
- *Are the models taking into account the same actors?*
  - Yes: docking, interoperation, or integration
  - No: collaboration or interoperation

[1] See http://www2.cesm.ucar.edu.

[2] See http://www.globalchange.umd.edu/models/gcam.

- *Are the models considering the same spatial region in the same timeframe?*
  - Yes: docking, interoperation, or integration
  - No: collaboration or interoperation

REQUIREMENTS FOR MULTIMODELING TECHNOLOGY

A toolkit to support the development and use of complex system models in a rapid, flexible, and robust manner has seven basic requirements: extensibility, interoperability, common and extensible ontological framework, common interchange language, data management, scalability, and robustness. These requirements are discussed below.

Extensibility

To help analysts answer a diverse and changing set of questions, modeling toolkits need to be easily extensible so that already integrated tools can be refined and new measures and techniques can be added. Building toolkits as modular suites of independent packages enables more distributed software development and allows independent developers to create their own tools and then connect them to a growing federation of toolkits. Such an integrated system can be facilitated by the use of common visualization and analysis tools and a common interchange language, meeting common interoperable requirements, and providing individual programs as web services, or at a minimum specifying their input and output in XML or a common interchange format. A number of initiatives to move various types of tools and models in a toolkit direction are currently underway at the U.S. Department of Defense, the Defense Advanced Research Projects Agency, and the National Science Foundation.

Interoperability

All tools embedded in a toolchain need to be capable of reading and writing the same data formats and data sets. This capability would allow output from one tool to be used as input to other tools. Each tool does not have to be able to use all the data from the other tools, but it does have to be capable of operating on relevant subsets of data without altering the data format. While not a hard requirement for a toolchain, it would be beneficial to agree upon a common interchange language for input and output to ease the concatenation of existing tools created by various developers. Other key aspects of interoperability are the use of a common ontology for describing data elements, and the ability for tools to be called by other tools through scripts.

Ontologies

Models of complex systems seek to capture the primary entities or components of the system, the relationships between those entities, and the attributes of the entities and relationships. Ideally, the classes representing entities, relationships, and attributes will form an ontology for representing complex system data. Tools are needed to automatically derive ontologies from the data or theory.

For relational data, most of the focus has been on connections between agents (social networks). This approach needs to be expanded for three reasons. First, multilink data sets have recently become available, but there has been little success in defining ontologies with multiple entity classes. Second, most existing tools are built on the assumption that there is only a single relation type at a time. Third, few data sets contain attributes, and so there has been little attention to defining appropriate attribute classes. As a result, there are relatively few candidate ontologies from which to choose.

Interchange Language

A common data interchange language ensures the consistent and compatible representation of various networks or identical networks at various states, and also facilitates data sharing and fusion. Therefore, network data collected using various techniques or by different people, stored and maintained in different databases with different structures, and used as input and output of various tools need to be represented in a common format. A common format or data interchange language—engineered for flexibility and compatibility with a variety of tools—can enable different groups to run the same tools and share results, even when the input data cannot be shared. Interchange language requirements exist for various classes of models, for example DyNetML for high-dimensional network models (Tsvetovat et al., 2003) or graphML for simple networks (Brandes et al., 2013), but there are no general requirements for complex system models.

Data Storage and Management

If multiple models that receive and return data are used in a project, a data storage and management system, typically a database, is needed. A database function that enables information to be added to networks that are stored in a database and then analyzed can lead to a more complete picture of social systems. Moreover, Structured Query Language (SQL)-type databases include tools for data search, selection, and refinement. Because there is no common database structure in the intelligence domain, translation and management tools are needed to combine data across data sets and convert between interchange languages. A key requirement here is a common ontology (as previously noted) so that diverse structures can be utilized and data can be rapidly fused together. Another difficulty is that most of the currently available relational data is contained in raw text files or stored in excel files, which complicates augmenting relational data with attributes, such as a person's age or gender. Utilization of SQL databases instead of these other formats will facilitate the handling of relational data.

Scalability of Results

Many models are designed and tested with small data sets, or are used to generate a small number of results. In general, however, these models should be analyzed from a response surface perspective. A response surface in a model is formed by the relationship between one or more explanatory variables and one or more response variables (i.e., the relation of inputs to outputs; Box and Wilson, 1951). For simulations, a good virtual experiment should be conducted in which the analyst systematically collects information on the response surface. When the number of variables is small, designs such as described by Box and Behnken (1960) can be used. However, as the number of variables increases, more exploratory techniques, such as active nonlinear tests (Miller, 1998), are needed. Once the virtual experiment has been completed, the results are analyzed and the best-fitting nonlinear model is estimated.

Tools supporting response surface methodology are available in many statistical and mathematical packages such as MATLAB. Analyzing models in this manner increases dramatically the number of replications needed and the overall size of the virtual experiment. Substantial work is typically required to make models sufficiently scalable to generate the response surface in a reasonable amount of time. To overcome the standard problem in response surface analysis of simulation results, the Tera-Grid (Catlett et al., 2007) can be used to generate a sufficiently large population of data to obtain sufficient variation in underlying parameters and ensure robustness of results.

Robustness

Models need to be robust in the face of missing data and common data errors. Key aspects of robustness include the following: (1) measures should be relatively insensitive to slight modifications of the data, and (2) the tools should be able to be run on data sets with diverse types of errors and varying levels of missing data.

Appendix B

Computation

EVOLUTION OF COMPUTING

In 1946, John von Neumann created the first direct ancestor of modern digital computing (Dahan-Dalmedico, 2001). In early computers, instructions and data were sent to a central processing unit, where single operations on every clock cycle acted on their operands held in registers, and the results were sent back to memory. As processors got faster reliably, computing power grew following a power-law relationship. A major technological upheaval occurred with the advent of vector processing in the 1970s. Vector processors allowed a steady data stream to flow through a pipeline of operations in sequence so that many operations could be performed concurrently. This approach led to specialized machines that were dramatically faster and opened up revolutionary new computational capabilities. Owing to specialized hardware, such as memory technologies capable of continuously feeding the vector registers, these machines were expensive, and so computations per dollar remained on the same power-law curve.

This power-law behavior is often referred to as Moore's law, based on Gordon Moore's observation that transistor density on a chip doubled roughly every 18 months (Chien and Karamcheti, 2013; Kogge et al., 2008). Another feature of microprocessor design, known as Dennard scaling, enabled the power density on a chip to remain constant as transistors became smaller. The results of Moore's law and Dennard scaling allowed the explosive growth in computational capacity to continue for several decades and made possible the personal computing revolution of the 1980s. Computing became inexpensive and ubiquitous, and a mass market for commodity parts developed. As a result, systems for parallel computation could be built using commodity parts, rather than the specialized hardware of vector supercomputers. Today, nearly all supercomputers are built from parts destined for the mass market.

Current computing technologies are reaching some physical limits (see Kogge et al., 2008, for an indispensable guide to hardware trends of the next decade). In particular, Dennard scaling no longer holds for conventional silicon technologies today. This is causing the electrical power needs of faster chips to grow faster than the mass-market applications can support. Thus, raw processing speed of a single arithmetic operation has stalled at about 1 GHz now for a decade. All advances in computational volume are delivered by packing ever more concurrency onto a chip. As outlined by Kogge et al. (2008), this yields systems with millions of commodity parts, and so high rates of hardware failure must be taken into account in system design.

Simultaneously, the mass market for computing itself is now driven by the mobile market, rather than desktops and servers. This has even more stringent power requirements than the first generation of commodity computing,

and it is likely that this market is already economically dominant. Figure B.1 shows how the difference in power-law slopes drove vectors out of the market, and why this is likely to be repeated as mobile chips replace today's high-performance computing processors.

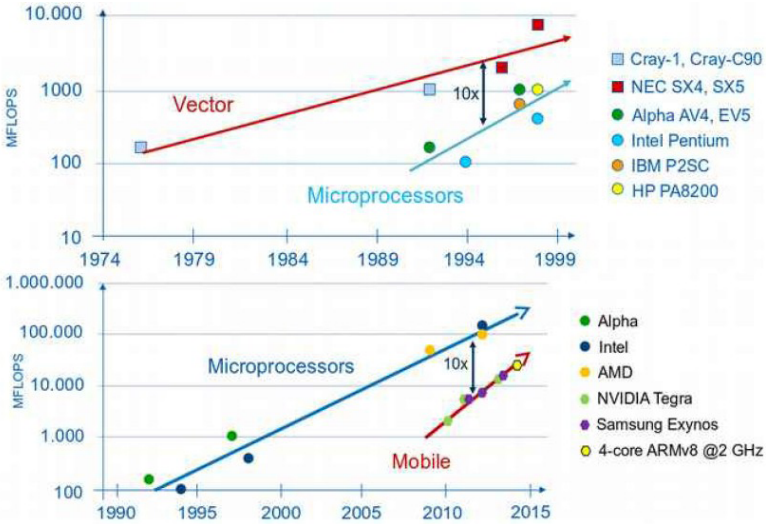


FIGURE B.1 The inexorable triumph of commodity computing over specialized computing. SOURCE: http://www.nextplatform.com/2015/06/16/mont-blanc-sets-the-stage-for-arm-hpc/. © Barcelona Supercomputing Center or BSC-CNS.

MODERN HIGH-PERFORMANCE COMPUTING

Below is a survey of hardware and software for the current generation of high-performance computing technologies for computation- and data-intensive applications.

Computation

The basic conceptual model of a processor consists of (a) arithmetic-logic units; (b) registers, which contain data to which arithmetic-logic units have read and write access; and (c) memory, where results are written. The finite representation of real numbers uses floating-point arithmetic. Floating-point arithmetic is nonassociative (i.e., $a+[b+c]$ may not be equal to $[a+b]+c$), and so the results are dependent on the order of operations. The memory location of a particular word is its address. At the end of a program, results held temporarily in memory are written to storage.

The increase in computing speed made possible by the miniaturization of complementary metal oxide semiconductor technology is coming to an end, and speed increases now come from executing many instructions in parallel. When two concurrent computations use the same memory address, there is danger of a race condition, in which the results depend on the order of completion of operations. The central problem of parallel computing is running many operations in parallel, while avoiding race conditions and minimizing synchronization costs.

Parallel computing has two basic modes: single instruction, multiple data (SIMD) and multiple instruction, multiple data (MIMD). In SIMD parallelism, the same computation is performed on multiple data streams. SIMD architectures include the graphical processing unit, which was primarily designed for video but is now used for other applications, and pipelining, which includes vector processing. In pipelining, the data flow through a sequence

of instructions, and one result is delivered every clock cycle. Vector machines have specialized vector registers for this purpose. The Cray-1 had 64 vector registers (known as the vector length, measured in 64-bit words) and later generations of vector machines (e.g., NEC-SX 9) have vector lengths up to 256. Some modern conventional processors contain vector instructions, but with a much smaller vector length, usually 4 or 8.

In MIMD parallelism, completely different instructions are executed in parallel. This can take the form of task parallelism, in which the computation is composed into a task graph with dependencies, and independent tasks are scheduled in parallel.

Memory and Storage

In early parallel computing, different processors accessed the same memory banks. However, there are physical limits on the number of connections that can be made to a single memory bus, typically said to be about 16. For higher numbers, many shared-memory computers, each with its own memory, are networked into a cluster. This distributed memory computing architecture is often called symmetric multiprocessor or, more commonly, commodity clusters because the processors used are generally mass market and unspecialized. A recent processor architecture, called many-integrated core, puts an entire symmetric multiprocessor on a chip. Intel's Xeon-Phi line (e.g., Knights Landing) is based on a many-integrated core architecture composed of individual Xeon (x86) cores.

Because memory access is slow compared to computation, there are layers of intermediate storage called caches. Current technologies often include many levels of cache (labeled L1, L2, etc., with L1 being closest to the registers). Each level of cache operates at a different speed and has a different size and a different unit (number of words simultaneously written) of access, called a cache line.

Storage is generally based on magnetic spinning disk technology. Parallelism in storage allows striping across multiple disks, which overcomes limits on the speed of writing to a single disk on a spindle. Parallel file systems (e.g., Lustre, GPFS, and Panasas) handle this task transparently. Tapes are often used for archival storage. Because tape is a linear medium (i.e., the tape needs to be spooled to a particular data location), it is not suitable for random access. Hierarchical storage management technology is used to handle many levels of storage: nearline, farline, and even offline.

New technologies are blurring the distinction between memory and storage. Nonvolatile random access memory and new hardware features, such as multichannel dynamic random access memory in Xeon-Phi (which can be configured either as a storage buffer or L4 cache), provide a continuous, many-level hierarchy spanning registers to tape. Programming such a hierarchy for computation- and data-intensive applications is one of the greatest high-performance computing challenges of our times.

Software

Concurrent streams of execution that access shared memory are called threads. A standard programming model for thread-based execution is OpenMP, which places directives or pragmas inside code indicating that a particular code section allows thread-safe parallel execution. Parallel execution streams in distributed memory are called processes, and the standard programming model is the message-passing interface (MPI), which allows communication and synchronization between processes. Hybrid MPI/OpenMP codes are currently common practice in compute-intensive applications. Partitioned global address space languages allow distributed memory to be programmed as though it were a unified global address space. Such languages include Co-Array Fortran, Unified Parallel C, Titanium (Java), and specialized languages not in wide use, such as Chapel and X10.

A similar approach is emerging for parallel data-intensive applications. MapReduce is a standard programming model for such applications. A computation, such as a search query, is dispatched to multiple data streams in parallel in the Map step, and the results are collated in the Reduce step. File systems designed specifically for

MapReduce, such as the Hadoop distributed file system, are an emerging alternative to the standard POSIX-based file systems of hierarchical directories and files common on commodity platforms.

Appendix C

Biographical Sketches of Committee Members

David M. Higdon, Chair, is a professor in the Social Decision Analytics Laboratory at Virginia Tech. Previously, he spent 10 years as a scientist or group leader of the Statistical Sciences Group at Los Alamos National Laboratory. Dr. Higdon holds a B.A. and an M.A. in mathematics from the University of California, San Diego, and a Ph.D. in statistics from the University of Washington. He is an expert in Bayesian statistical modeling of environmental and physical systems, combining physical observations with computer simulation models for prediction and inference. His research interests include space-time modeling; inverse problems in hydrology and imaging; statistical modeling in ecology, environmental science, and biology; multiscale models; parallel processing in posterior exploration; statistical computing; and Monte Carlo and simulation-based methods. Dr. Higdon has served on several advisory groups concerned with statistical modeling and uncertainty quantification, and co-chaired the National Research Council (NRC) Committee on Mathematical Foundations of Validation, Verification, and Uncertainty Quantification. He is a fellow of the American Statistical Association.

Robert L. Axtell is a professor and chair of the Department of Computational Social Science at George Mason University. Previously he was a senior fellow in the Economic Studies and Governance Studies programs at the Brookings Institution. He holds a B.S. from the University of Detroit and an interdisciplinary Ph.D. from Carnegie Mellon University, where he studied economics, computer science, and public policy. Dr. Axtell's research involves agent-based computational models of social phenomena, in which autonomous software agents—each agent representing an individual person—interact according to simple rules of behavior, with patterns and structure emerging at the aggregate level. His current focus is on the creation of entire artificial economies consisting of hundreds of millions of agents. His book *Growing Artificial Societies: Social Science from the Bottom Up* (MIT Press, 1996), co-authored with J.M. Epstein, is widely cited as an early statement of the potential of multiagent systems to more fully represent social processes.

Venkatramani Balaji is head of the Modeling Systems Group at the Geophysical Fluid Dynamics Laboratory and Princeton University. His group provides a software environment where scientific groups can develop new physics and new algorithms concurrently, and also coordinate their efforts. Dr. Balaji received an M.S. from the Indian Institute of Technology, Kanpur, and a Ph.D. from The Ohio State University, both in physics. He is an expert in parallel computing and scientific infrastructure, and has pioneered the use of model frameworks, such as the Flexible Modeling System, and community standards needed to construct climate models from independently

developed components sharing a technical architecture. Dr. Balaji served on the NRC Committee on a National Strategy for Advancing Climate Modeling. He is a sought-after speaker and lecturer and is committed to provide training in the use of climate models in developing nations, leading workshops to advanced students and researchers in South Africa and India.

Lawrence E. Buja directs the Climate Science and Applications Program at the National Center for Atmospheric Research (NCAR). The program examines societal vulnerability, impacts, and adaptation to climate change using climate change scenarios; vulnerability analyses; integrated analyses of changes in climate, land use, conventional pollution, biodiversity, and human systems; and decision support tools. Previously, Dr. Buja was scientific project manager for NCAR's Community Climate System Model, which simulates Earth's past, present, and future climates and is one of the models used by the Intergovernmental Panel on Climate Change (IPCC). Dr. Buja was a contributing author to both the third and fourth IPCC assessments. Dr. Buja also works with the World Bank, the Inter-American Development Bank, and other international agencies, applying NCAR's climate and social science expertise to help guide sustainable development strategies throughout the developing world. He has a B.S. and an M.S. in atmospheric science from Iowa State University and a Ph.D. in meteorology from the University of Utah.

Katherine V. Calvin is a scientist at the Pacific Northwest National Laboratory's Joint Global Change Research Institute. Prior to joining the institute in 2008, she spent 2 years as an international energy analyst at the U.S. Energy Information Administration. Dr. Calvin earned bachelor's degrees in mathematics and in computer science from the University of Maryland, and M.S. and Ph.D. degrees in management science and engineering from Stanford University. Her work focuses on model development and scenario analysis using a global change integrated assessment model, with an emphasis on assessing climate change impacts and potential adaptation, and examining the effects of bioenergy and land policy on land use. She also coordinates regional model comparison exercises of Asia and Latin America. Dr. Calvin was a lead author for the National Climate Assessment and a contributing author for an IPCC assessment, and is currently on the scientific steering committee for the Land Use Model Intercomparison Project.

Kathleen M. Carley is a professor of computation, organization, and society at the Institute for Software Research at Carnegie Mellon University. She also directs the university's Center for Computational Analysis of Social and Organizational Systems, which brings together network analysis, computer science, and organization science. Dr. Carley holds bachelor's degrees in economics and in political science from the Massachusetts Institute of Technology, and a Ph.D. in sociology from Harvard University. She uses organization theory, dynamic network analysis, social networks, multiagent systems, and computational social science to examine how cognitive, social, technological, and institutional factors affect individual, team, social, and policy outcomes in areas ranging from public health to counterterrorism to cybersecurity. Dr. Carley has participated in several NRC studies on modeling and intelligence needs, including the Committee on Modeling and Simulation for Defense Transformation and the Committee on the Future U.S. Workforce for Geospatial Intelligence. She is a fellow of the Institute of Electrical and Electronics Engineers (IEEE), and received the lifetime achievement award from the Mathematical Sociology section of the American Sociological Association, and the Simmel award for advances in social networks and network science from the International Network for Social Network Analysis.

Rebecca Castaño is division technologist for the Mission Systems and Operations Division and the discipline area program manager for Applied Sciences at the Jet Propulsion Laboratory. She works to ensure that relevant new technologies are developed, matured, validated, and infused into instruments and missions. From 2002 to 2007, she was the Supervisor of the Machine Learning Systems Group, which uses learning algorithms, data mining, knowledge discovery, pattern recognition, and automated classification and clustering to carry out automated analyses of remote sensing data. Dr. Castaño received her B.S. in electrical and computer engineering from the University

of Iowa and her M.S. and Ph.D. in electrical and computer engineering from the University of Illinois, where she focused on computer vision. Dr. Castaño's research interests include machine learning, computer vision, and pattern recognition. She has spent the past 5 years advancing the state of the art in onboard science analysis methods, and has contributed to software operating on Earth orbiters and Mars rovers. She received NASA's Exceptional Engineering Achievement Medal in 2008.

Ronald R. Coifman (NAS) is the Phillips Professor of Math and Computer Science at Yale University. His research interests include nonlinear Fourier analysis, wavelet theory, singular integrals, numerical analysis and scattering theory, and real and complex analysis. He is also interested in new mathematical tools for efficient computation and transcriptions of physical data, as well as their applications to numerical analysis, feature extraction recognition, and denoising. He is currently developing analysis tools for spectrometric diagnostics and hyperspectral imaging. Dr. Coifman served on the National Academies of Sciences, Engineering, and Medicine's Board on Mathematical Sciences and Their Applications and the NRC Committee on the Analysis of Massive Data. He is a recipient of the 1996 Defense Advanced Research Projects Agency Sustained Excellence Award, the 1996 Connecticut Science Medal, the 1999 Pioneer Award of the International Society for Industrial and Applied Science, and the 1999 National Medal of Science. Dr. Coifman is a member of the American Academy of Arts and Sciences, the Connecticut Academy of Science and Engineering, and the National Academy of Sciences. He received his Ph.D. from the University of Geneva.

Omar Ghattas is the John A. and Katherine G. Jackson Chair in Computational Geosciences, a professor of geological sciences and mechanical engineering, and director of the Center for Computational Geosciences at The University of Texas at Austin. Previously, he was a professor at Carnegie Mellon University for 16 years. Dr. Ghattas earned a B.S. in civil engineering and M.S. and Ph.D. degrees in computational mechanics, all from Duke University. He has research interests in simulation and modeling of complex geophysical, mechanical, and biological systems on supercomputers, with specific interest in inverse problems and associated uncertainty quantification for large-scale systems. He served on several advisory groups on these topics, including the NRC Committee on Mathematical Foundations of Validation, Verification, and Uncertainty Quantification and the National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges. Dr. Ghattas is a recipient of the 2003 IEEE/Association for Computing Machinery Gordon Bell Prize for Special Accomplishment in Supercomputing, and is a fellow of the Society for Industrial and Applied Mathematics.

James A. Hansen is the head of the Meteorological Applications Development Branch in the Marine Meteorology Division of the Naval Research Laboratory. Previously, he was an associate professor in the Earth, Atmospheric, and Planetary Sciences Department at the Massachusetts Institute of Technology. Dr. Hansen has a B.S. and an M.S. in aerospace engineering from the University of Colorado. He received a Rhodes Scholarship and obtained his Ph.D. in atmospheric, oceanic, and planetary physics from Oxford University. His research interests focus on the estimation of environmental forecast uncertainty and the use of that uncertainty in decision making. One of his recent projects was the development of the Pirate Attack Risk Surface product, which uses intelligence, meteorological, oceanographic, and adversarial behavioral information to estimate and communicate the risk of pirate attack. Dr. Hansen served on the NRC Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts.

Anna M. Michalak is a faculty member in the Department of Global Ecology at the Carnegie Institution for Science. Previously, she was an associate professor at the University of Michigan. Dr. Michalak holds a B.S. in environmental engineering from the University of Guelph, Ontario, and an M.S. and a Ph.D. in civil and environmental engineering from Stanford University. Her research interests focus on characterizing complexity and quantifying uncertainty in environmental systems with the goal of improving our understanding of these systems

and our ability to forecast their variability. Application areas include estimating greenhouse gas emissions and sequestration, understanding linkages between climate variability and water quality, and characterizing the Earth system. A common theme of her research is the development and application of statistical and geostatistical data fusion methods for optimizing the use of limited in situ and remote sensing environmental data. Dr. Michalak has served on several scientific committees and organized workshops on simulating complex systems, and on observations and models for inferring greenhouse gas emissions. She is a recipient of the Presidential Early Career Award for Scientists and Engineers.

Shashi Shekhar is the McKnight Distinguished University Professor in the Department of Computer Science at the University of Minnesota. He holds a B.Tech. in computer science from the Indian Institute of Technology in Kanpur, India, and an M.S. and a Ph.D. in computer science from the University of California, Berkeley. Dr. Shekhar pioneered the research area of spatial data mining via pattern families (e.g., collocation and spatial outliers) and also has research interests in spatial databases. He co-authored a textbook on spatial databases and received IEEE's Technical Achievement Award for contributions to spatial databases, data mining, and geographic information systems. Dr. Shekhar has served on two NRC committees concerned with geospatial intelligence—the Committee on Basic and Applied Research Priorities in Geospatial Science for the National Geospatial-Intelligence Agency and the Committee on Future Workforce for Geospatial Intelligence. He is also a member of the Computing Community Consortium Council and chaired its spatial computing workshop in 2014. He is a fellow of the American Association for the Advancement of Science and of IEEE.

Shaowen Wang is a professor of geography and geographic information science and a faculty affiliate in the Department of Computer Science, Department of Urban and Regional Planning, and School of Information Sciences at the University of Illinois at Urbana-Champaign. He is also associate director for cyberGIS at the National Center for Supercomputing Applications, and founding director of the university's CyberGIS Center for Advanced Digital and Spatial Studies. Dr. Wang received a B.S. in computer engineering from Tianjin University, an M.S. in geography from Peking University, and an M.S. of computer science and a Ph.D. in geography from the University of Iowa. His research interests include geographic information science and systems, advanced cyberinfrastructure and cyberGIS, complex environmental and geospatial problems, computational and data sciences, high-performance parallel and distributed computing, and spatial analysis and modeling. He is president of the University Consortium for Geographic Information Science, and served on its board of directors from 2009 to 2012. He also serves on the advisory board of the National Science Foundation's Extreme Science and Engineering Discovery Environment program. He was an NCSA Fellow in 2007, and received the NSF CAREER Award in 2009.

Appendix D

Acronyms and Abbreviations

CMIP	Coupled Model Intercomparison Project
DoD	U.S. Department of Defense
FLOPS	floating-point operations per second
GIS	geographic information system
IPCC	Intergovernmental Panel on Climate Change
ISIS	Islamic State in Iraq and Syria
MCMC	Markov chain Monte Carlo
MIMD	multiple instruction, multiple data
MPI	message-passing interface
NGA	National Geospatial-Intelligence Agency
SIMD	single instruction, multiple data
SQL	Structured Query Language