TERRA POPULUS

A Global Population/Environment Data Network

Steven Ruggles, Catherine A. Fitch, Jonathan Foley, Steven M. Manson, and Matthew Sobek
University of Minnesota
September 23, 2011

ABSTRACT

Terra Populus, part of NSF's new DataNet initiative, will develop organizational and technical infrastructure to integrate, preserve, and disseminate data describing changes in the human population and environment over time. A plethora of high-quality environmental and population datasets are available, but they are widely dispersed, have incompatible or inadequate metadata, and have incompatible geographic identifiers. The new infrastructure will enable researchers to identify and merge data from heterogeneous sources to study the relationships between human behavior and the natural world. *Terra Populus* will partner with data archives, data producers, and data users to create a sustainable international organization that will guarantee preservation and access over multiple decades.

Terra Populus: A Global Population/Environment Data Network (TerraPop) is a new data infrastructure project funded by the National Science Foundation Office of Cyberinfrastructure as part of the DataNet initiative.[1] TerraPop will develop organizational and technical infrastructure to integrate, preserve, and disseminate data describing population and environment on a global scale over the past two centuries, including data on human population characteristics, land use, land cover, and climate change. It will make these data interoperable across time and space, disseminate them to the public and to multiple research communities, and preserve these precious resources for future generations. More broadly, TerraPop will be a model for the sustainable expansion, maintenance, and improvement of a global data resource.

Population growth and Environmental Change

Over the past five decades, the world's population more than doubled. Sharp interregional differences in growth rates—together with unprecedented urbanization and international migration—led to dramatic spatial redistribution of population. Economic changes were equally remarkable. World per-capita gross domestic product roughly doubled, but that expansion was uneven, marked by growing inequality in many regions and little convergence in economic development between rich and poor countries [1-4].

The rate of population growth during the past half century was unprecedented and is unlikely to recur. In virtually every country, fertility rates are declining. This is creating another massive structural change: a shift in the age composition of the world's population, which will strain social, economic, and environmental resources as twentieth-century birth cohorts enter old age [5, 6]. Other dramatic demographic trends—rising urbanization and international migration, industrialization of the developing world, and improvements in education and health—are likely to continue or accelerate in coming decades.

The extraordinary levels of global demographic and economic growth since the 1950s have had ominous consequences: alarming environmental degradation, resource depletion, and climate change [2, 3]. In just the last 50 years, food and water consumption roughly tripled, alongside a four-fold increase in use of fossil fuels. There is rapidly increasing pressure on global land resources, biodiversity, and "ecosystem goods and services" [7, 8]. The average global temperature has gone up 0.74˚Celsius over the past century, and is now rising at an accelerating pace; predictions for temperature increase over the next century range from 1.1˚ to 6.4˚. Sea levels are rising, and the oceans are growing more acidic. New precipitation patterns—including increased precipitation in high latitudes and decreased rainfall in subtropical regions—are becoming more pronounced [9, 29]. Deforestation and pollution are compounding the direct effects of global warming and contributing to the destruction of ecosystems and decline of biodiversity [8, 9].

Changes in population size, characteristics, and behavior lie at the heart of these environmental challenges. The key drivers of change—especially fossil fuel emissions and deforestation—are direct consequences of population growth and economic development. Conversely, environmental change has profound implications for demographic behavior. Flooding, erosion of coastal areas, destruction of ecosystems, and drought and degradation of water supplies at lower latitudes all have consequences for human populations, such as mass migration, food scarcity, and increased armed conflict.

Our understanding of the interactions of population and environment has been hampered by the dearth of internationally comparable data. Newly-available population data closely integrated with data on the environment will allow us to describe the unfolding transformation of human

[1] http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

and ecological systems. Data on the human population are crucial for understanding changes in the Earth's biological and climate processes; equally important, data on climate and land provide essential tools for understanding the impact of environmental change on human behavior. By creating a framework for locating, analyzing, and visualizing the world's population and environment in time and space, TerraPop will provide unprecedented opportunities for investigating the agents of change, assessing their implications for human society and the environment, and developing policies to meet future challenges.

The National Research Council [5] makes a compelling case for the development and use of cross-national and cross-temporal data sources, declaring that "national and international funding agencies should establish mechanisms that facilitate the harmonization of data collected in different countries." The report argues that "cross national studies conducted within a framework of comparable measurement can be a substantially more useful tool for policy analysis than studies of single countries." The National Science Foundation (NSF) Cyberinfrastructure Council describes "a vision in which science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved" [10]. Scientific advances and policy insights are greatest when users with varying theoretical perspectives and models have access to the same data. NSF's 2009 report on *Solving the Puzzle* of climate change reinforces the importance of cyberinfrastructure for international data sharing and collaboration [29]. The TerraPop infrastructure directly addresses these needs. It will preserve and integrate social and natural science data and metadata from across the globe, disseminate them across multiple research communities, and translate these digital data into usable knowledge.

TerraPop Goals

TerraPop will create generalized methods and systems for integrating and disseminating spatiotemporal data. Our initial work will focus on four specific kinds of data: (1) census and survey microdata describing the characteristics of individuals and their families and households; (2) aggregate census and survey data, describing the characteristics of places, including aggregate population characteristics, land use, and land cover; (3) remote-sensing data describing land cover and other environmental characteristics; and (4) climate data describing temperature, precipitation, and other climate-related variables. All four data types have an important temporal dimension; most of the data span the past five decades, and many sources reach back to the nineteenth century.

In partnership with most of the world's national statistical agencies, as well as data archives and research centers, project collaborators have assembled the world's largest collection of spatiotemporal population data. This massive effort was financed with approximately $100 million contributed by funding agencies in North America and Europe [11-14]. TerraPop will merge these human population data with a vast body of environmental data derived from government land-use statistics, satellite imaging, and climate records. Project collaborators have already gathered census-based land-use and land-cover records for over 22,000 different political units around the world, and are extending the collection back as far as possible [15-18]. TerraPop will fuse these census-based records with major global land-cover databases derived from satellite imagery [19-21]. We will also incorporate historical data on climate—including station measurements of temperature, precipitation, and cloud cover and geospatially-gridded data products—which sometimes date back to the nineteenth century [22, 23, 24]. The climate data collection will leverage fundamental contributions of TerraPop collaborators to the improvement of historical data on precipitation and temperature [25-28], relying on wavelet analysis, singular spectrum analysis, and related techniques to fill spatiotemporal gaps in the source data.

By integrating the most comprehensive global data collections on land use and climate change with the major global collections of population data, TerraPop will constitute a unique international reference collection for investigating changes in the human-environment system. Our goal is to support analysis from local to global scales and from historical to contemporary epochs.

The central objective is to create a sustainable organization for the integration, preservation, and dissemination of global-scale data on the human population and the environment over a decades-long time frame. The project involves four key elements: (1) developing an archive, (2) creating new software and methods for data access and integration, (3) conducting education and outreach; and (4) designing a new organizational structure. The four elements of the project are not sequential phases; work will occur on each element of the project simultaneously. We will produce a "prototype" in the first 18 months that incorporates components of each project element.

1. Archival development. We must develop accession, preservation, and protection policies and procedures. We will collect and preserve high-value datasets describing the world's population and environment over the past two centuries. We will develop interoperable metadata to describe those datasets.

2. Data dissemination and analysis. We must develop tools and procedures to manage the collections, make them interoperable, and disseminate them to the research community

3. Education and outreach. We will develop a training program and collaborate with private-sector and nonprofit organizations to engage the scientific community and the public and reach the broadest possible audience.

4. Organizational structure. We will collaborate with data archives, data producers, and data users to design and implement institutional structures and policies that will provide for long-run governance and financial sustainability and which will guarantee preservation and access over multiple decades.

Deliverables

The following sections outline key deliverables for each of the four major project elements**.**

1. Archival development. We will establish a committee to develop a data accession plan that will allow us to prioritize data ingest. We will establish a committee to develop data preservation and protection policies, and we will implement those policies. We will acquire datasets and develop metadata to describe them and allow data integration. This element includes the following deliverables:

1.1 Ingest and Preservation

- Develop data accession and deaccession policies.
- Collect and preserve selected datasets relating to population and the environment.
- Inventory and evaluate land-use, climate, and small-area census data for inclusion in the infrastructure. Design and implement a distributed preservation and replication structure.
- Develop data protection policies and procedures.

1.2 Integration and Interoperability

- Implement spatial algorithms to simplify data fusion.
- Improve metadata describing geographic units identified by statistical agencies around the world in different periods to allow data integration across time and between sources.
- Account for boundary changes through re-aggregation or interpolation to allow analysis of change over time.
- Carry out variable-level data integration for interoperability of statistical data from heterogeneous sources on comparable topics across time, space, and type of data.
- Create integrated variable coding across datasets within data classes (population microdata, population aggregate data, land use, land cover, climate) to maximize comparability over time and between countries.
- Create compatible variables across aggregate population data and population microdata.

1.3 Metadata development

- Develop or improve machine-actionable metadata for all data collections.
- Design and implement tools and procedures to streamline metadata creation and management.

Element 2. Dissemination and Analysis. We will develop and implement software and methods to provide for access, analysis, and visualization of the datasets. These tools are needed to manage the collections, make them interoperable, and disseminate them to the research community. This element includes the following deliverables:

2.1 Data Access System

- Develop an open-source electronic dissemination application for census and survey microdata, aggregate census data, land use data, land cover data, and climate data.
- Build flexible data manipulation tools for merging data from diverse sources and multiple levels of analysis based on spatial location, even when the original geocoding is incompatible
- Build tools to allow users to specify the methods and scales used for geographic linking across datasets.
- Implement innovative dissemination methods that enhance search and discovery of information by diverse users.
- Develop an Application Programming Interface to allow interoperability with remote systems.

Figure 3 illustrates the architecture of the proposed dissemination system. The initial system will deliver merged microdata and environmental data in a range of formats suited to different analytic strategies. Planned enhancements of the system are shown with dashed lines.

2.2 Analysis, Visualization, and Community Tools

- Develop and implement high-speed microdata aggregation tools allowing users to construct customized population measures for small areas in real time directly from microdata.
- Implement a column-oriented storage layout that vertically partitions large data tables into smaller ones.
- Develop mapping and visualization tools oriented to general audiences.
- Design systems to foster scientific and educational collaborations across institutions, promote sharing of tools and knowledge, and reduce redundant effort, through wiki-enabled documentation, specialized forums that encourage collaborations among researchers, and searchable repositories for sharing software code.
- Create linkages between primary data, data extracts, bibliographic references, and relevant documents that are managed in the TerraPop repository.

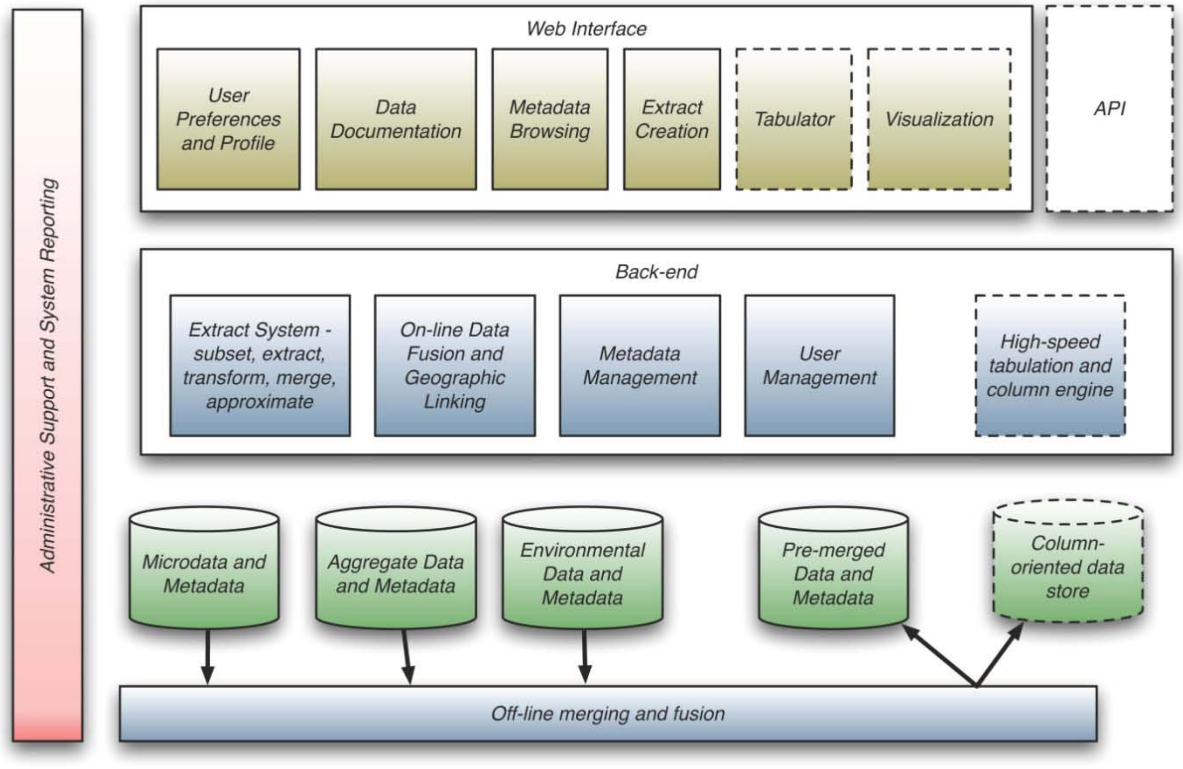
Administrative Support and System Reporting
Web Interface
User Preferences and Profile
Data Documentation
Metadata Browsing
Extract Creation
Tabulator
Visualization
API
Back-end
Extract System - subset, extract, transform, merge, approximate
On-line Data Fusion and Geographic Linking
Metadata Management
User Management
High-speed tabulation and column engine
Microdata and Metadata
Aggregate Data and Metadata
Environmental Data and Metadata
Pre-merged Data and Metadata
Column-oriented data store
Off-line merging and fusion

Figure 3. Data Access Architecture

Element 3. Education and Outreach. We will develop a training program and collaborate with private-sector and nonprofit organizations to engage the scientific community and the public and reach the broadest possible audience.

3.1 Education and Training

- Train and mentor graduate students and postdoctoral research associates working on the project to develop the next generation of scientific data experts.
- Develop a curriculum of web-based training focusing on TerraPop dissemination and analysis tools. These self-study tutorials will be discrete modules that users can draw on for more information about a particular aspect of the data or analysis tools.

3.2 Outreach

- Conduct workshops and exhibits at conferences to inform diverse scientific and educational audiences of the availability of the resource.
- Partner with collaborators in the private sector and non-profit world to bring TerraPop to a broader public.

Element 4: Organizational Development. In collaboration with leading stakeholders around the world, the project will develop and implement a plan for a new permanent and self-sustaining data organization to provide preservation and access over multiple decades. This element includes the following deliverables:

- Form an international sustainability committee to negotiate and consult with key stakeholders, including data producers, data archives, libraries, researchers in multiple disciplines, educators, and the lay public.
- Produce a document describing the major options for governance and long-run sustainability within 18 months.
- By the end of year 3, the governance and sustainability plan will be finalized, and by the end of year 5, it will be implemented.
- Acquire and implement technology to assure communications across collaborators.
- Develop and implement quantitative and qualitative measures assessing TerraPop success. Basic criteria in this assessment will include the quantity of data distributed; the number of data users, broken down by type of user (e.g., scientists by discipline, students by academic level, and so on); and indicators of participation in the TerraPop web communities.

Development Strategy

TerraPop has both short-range and long-range goals. In the short run, we must implement innovative software and metadata to allow interoperability of major collections of data relating to population and the environment. The Minnesota Population Center has experience addressing the risks associated with software and infrastructure development. In the long run, we must create a viable organization that can provide data access and preservation over a decades-long period.

The long-run TerraPop development strategy is designed to answer three interconnected questions: (1) Organizational sustainability: how can we develop institutional structures that ensure wise and efficient leadership and management of the network over the long run? (2) Financial sustainability: how can we be assured of sufficient funding to maintain and improve the data collection and maintain the network over a decades-long time-frame? (3) Technological sustainability: how can we adapt to changing technology and meet ever increasing user demand for services?

Two decades ago, much of the infrastructure we propose to build would have been inconceivable. Because of changing standards—especially in the realm of metadata—much data infrastructure from even a decade ago is already obsolete. Accordingly, TerraPop must be prepared to continuously adapt to a shifting technological environment. We must also respond to evolving user needs. The entire approach to the use of technology for research and education is likely to shift in coming decades. There will be new ways of interacting with data, and users will want to repurpose existing data collections to serve new needs. We cannot predict what technological innovations and changes in user needs and expectations will occur over the coming two, three, or four decades. We do not know what we will have to build to respond to change, and we certainly do not know how much it will cost. We can, however, be confident that technological change will occur. Institutions responsible for data must be agile enough to keep the data collections viable and to capitalize on new opportunities. If we can create an organization that has both sufficient money and effective leadership, the third challenge—adapting to the shifting technological environment—will be solvable. Accordingly, the greatest sustainability challenge is not technology, but organizational development.

Our goal is to form an organization that maximizes long-run financial security; guarantees continuity of creative, efficient, and responsive leadership over the very long run; and ensures that all stakeholders—data producers, data archivists and librarians, researchers and educators, and the lay public—are engaged in setting organizational priorities and invested in the success of the organization. None of the existing models are ideal. Accordingly, we will design a new organizational structure that effectively balances these competing goals. We are guided by three general principles. First, to minimize risk, we need multiple and diverse sources of funding for the organization. Second, we need effective mechanisms to ensure the active involvement of a wide range of stakeholders in governance of the organization. Third, we need a large and financially-secure institutional parent that will assume ultimate responsibility for preserving the collection.

The best solution for TerraPop, we believe, is a hybrid organization that has some characteristics of an academic library and some characteristics of a membership organization. Over the next several years, in close consultation with collaborators and stakeholders around the world, we will design a robust organizational structure for TerraPop that will both ensure long run preservation and be adaptable and responsive to the needs of data users and producers.

References Cited

1. Ferreira, F. and M. Ravallion, Poverty and Inequality: The Global Context, in *The Oxford Handbook on Economic Inequality*, W. Salverda, B. Nolan, and T. Smeeding, Editors. 2009, Oxford University Press.
2. World Bank, World Development Indicators 2009: Washington D.C. http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:21725423~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html.
3. O'Neill, B., C.F. MacKellar, and W. Lutz, *Population and Climate Change*. 2001, Cambridge: Cambridge University Press.
4. Lutz, W. and A. Goujon, The World's Changing Human Capital Stock: Multi-State Population Projections by Educational Attainment. *Population and Development Review*, 2001. 27(2): 323-339. http://www3.interscience.wiley.com/journal/118986314/abstract?CRETRY=1&SRETRY=0.
5. National Research Council, *Preparing for an Aging World: The Case for Cross-National Research*. 2001, Washington D.C.: National Academy Press. http://www.nap.edu/openbook.php?isbn=0309074215.
6. United Nations. *World Population Prospects.* 2006; Available from: http://www.un.org/esa/population/publications/wpp2006/WPP2006_Highlights_rev.pdf. Accessed 5/14/09.
7. Foley, J.A., R. DeFries, G.P. Asner, C. Barford, G. Bonan, S.R. Carpenter, F.S. Chapin, M.T. Coe, G.C. Daily, H.K. Gibbs, J.H. Helkowski, T. Holloway, T. Howard, E.A. Howard, C.J. Kucharik, C. Monfreda, J.A. Patz, I.C. Prentice, N. Ramankutty, and P.K. Snyder, Global Consequences of Land Use*. Science*, 2005. 309: 570-574. http://www.sciencemag.org/cgi/content/full/309/5734/570.
8. Fischlin, A., G.F. Midgley, J.T. Price, R. Leemans, B. Gopal, C. Turley, M.D.A. Rounsevell, O.P. Dube, J. Tarazona, and A.A. Velichko, Ecosystems, Their Properties, Goods and Services, in *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden, and C.E. Hanson, Editors. 2007, Cambridge University Press, Cambridge. http://www.ipcc.ch/ipccreports/ar4-wg2.htm.
9. IPCC, Summary for Policymakers, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, Editors. 2007, Cambridge University Press Cambridge, United Kingdom and New York, NY, USA. http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-spm.pdf.
10. NSF Cyberinfrastructure Council. *Cyberinfrastructure Vision for 21st Century Discovery.* 2007; Available from: http://www.nsf.gov/pubs/2007/nsf0728/index.jsp. Accessed 5/10/09.
11. Ruggles, S., M. Sobek, J.T. Alexander, C. Fitch, R. Goeken, P.K. Hall, M. King, and C. Ronnander, Integrated Public Use Microdata Series: Version 4.0 [Machine-Readable Database]. 2008, Minnesota Population Center [producer and distributor], Minneapolis, Minnesota. http://usa.ipums.org/usa/.
12. Minnesota Population Center, Integrated Public Use Microdata Series–International: Version 5.0. 2009, Minneapolis: University of Minnesota. https://international.ipums.org/international/.
13. North Atlantic Population Project and Minnesota Population Center, (NAPP): Complete Count Microdata. Version 2.0 [Computer Files]. 2008, Minneapolis, MN: Minnesota Population Center [distributor]. http://www.nappdata.org.
14. Minnesota Population Center, National Historical Geographic Information System. 2009, Minneapolis, MN: University of Minnesota. http://www.nhgis.org/.

15. Ramankutty, N. and J.A. Foley, Characterizing Patterns of Global Land Use: An Analysis of Global Croplands Data. *Global Biogeochemical Cycles*, 1998. 12(4): 667-685. http://www.agu.org/pubs/crossref/1998/98GB02512.shtml.
16. Ramankutty, N. and J.A. Foley, Estimating Historical Changes in Global Land Cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles* 1999. 13(4): 997-1027. http://www.agu.org/pubs/crossref/1999/1999GB900046.shtml.
17. Ramankutty, N., A.T. Evan, C. Monfreda, and J.A. Foley, Farming the Planet: 1. Geographic Distribution of Global Agricultural Lands in the Year 2000. *Global Biogeochem. Cycles*, 2008. 22: GB1003, doi:10.1029/2007GB002952. http://www.agu.org/pubs/crossref/2008/2007GB002952.shtml.
18. Monfreda, C., N. Ramankutty, and J.A. Foley, Farming the Planet: 2. Geographic Distribution of Crop Areas, Yields, Physiological Types, and Net Primary Production in the Year 2000. *Global Biogeochemical Cycles*, 2008. 22: GB1022, doi:10.1029/2007/GB002947. http://www.agu.org/pubs/crossref/2008/2007GB002947.shtml.
19. *Modis. (Moderate Resolution Imaging Spectroradiometer).* Available from: http://modis.gsfc.nasa.gov/. Accessed 5/9/09.
20. *Landsat.* Available from: http://landsat.gsfc.nasa.gov/. Accessed 5/9/09.
21. *ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer).* Available from: http://asterweb.jpl.nasa.gov/. Accessed 5/9/09.
22. New, M.G., M. Hulme, and P.D. Jones., Representing Twentieth-Century Space-Time Climate Variability. Pt. 1. Development of a 1961–1990 Mean Monthly Terrestrial Climatology. *Journal of Climate*, 1999. 12: 829–56. http://ams.allenpress.com/archive/1520-0442/12/3/pdf/i1520-0442-12-3-829.pdf.
23. New, M.G., M. Hulme, and P.D. Jones., Representing Twentieth-Century Space-Time Climate Variability. Pt. 2. Development of a 1901–1996 Monthly Terrestrial Climate Field. *Journal of Climate* 2000. 13: 2217-38. http://badc.nerc.ac.uk/data/cru/cru05_doc.pdf.
24. New, M., D. Lister, M. Hulme, and I. Makin, A High-Resolution Data Set of Surface Climate over Global Land Areas*. Climate Research* 2002. 21. http://www.cru.uea.ac.uk/cru/data/tmc/new_et_al_10minute_climate_CR.pdf.
25. Narisma, G.T., J.A. Foley, R. Licker, and N. Ramankutty, Abrupt Changes in Rainfall During the Twentieth Century*. Geophysical Research Letters*, 2007. 34, L06710, doi:10.1029/2006GL028628. http://www.agu.org/pubs/crossref/2007/2006GL028628.shtml.
26. Botta, A., N. Ramankutty, and J.A. Foley, Long-Term Variations of Climate and Carbon Fluxes over the Amazon Basin*. Geophysical Research Letters* 2002. 29(9): 33-1 to 33-4, doi:10.1029/2001GL013607. http://www.agu.org/pubs/crossref/2002/2001GL013607.shtml.
27. Costa, M.H. and J.A. Foley, Trends in the Hydrologic Cycle of the Amazon Basin*. Journal of Geophysical Research (Atmospheres)* 1999. 104 (D12)(14): 189-14,198. http://www.agu.org/pubs/crossref/1999/1998JD200126.shtml.
28. Schlesinger, M.E., N. Ramankutty, and N. Andronova, Temperature Oscillations in the North Atlantic*. Science* 2000. 289 (5479): 547-548. http://www.sciencemag.org/cgi/content/full/sci;289/5479/547b?ck=nck.
29. National Science Foundation, Solving the Puzzle: Researching the Impacts of Climate Change Around the World. Released May 14, 2009. Available from: http://www.nsf.gov/news/special_reports/climate/pdf/NSF_Climate_Change_Report.pdf. Accessed 5/14/09.