A Model-Inspired Sampling Network Design and Representativeness Methodology for the Arctic

Forrest M. Hoffman[1], Jitendra Kumar[1], Stan D. Wullschleger[1], Larry D. Hinzman[2], and Edward A. G. Schuur[3]

[1]Oak Ridge National Laboratory, [2]University of Alaska-Fairbanks, [3]University of Florida

January 12, 2013

Arctic Observing Summit (AOS) 2013
Theme 2 – Observing System Design and Coordination

1 Executive Summary

Climate warming is projected to be especially pronounced at high latitudes. Rising temperatures and associated impacts at local to Pan-Arctic scales are likely to be significant. National and international networks that target spatially and temporally intensive monitoring of land, atmosphere, and ocean processes are critically important investments for understanding our changing climate and environmental responses to those changes. Despite the potential implications of climatic change to environmental, societal, and political systems, the Arctic has a limited record of observations, both in terms of record length and spatial coverage. Resource and logistical constraints limit the frequency, extent, and optimality of spatial sampling of environmental observations. In the Arctic, reduced access limits most station locations to low elevations near the coast or near to roads and population centers, necessitating the development of a systematic sampling strategy to maximize coverage and objectively represent environmental variability at scales that are important to the land manager, decision-maker, or other stakeholder. ***Described here is a quantitative methodology for designing observing networks, stratifying sampling domains, informing site selection, and determining the representativeness of measurement sites and networks.*** This analysis provides model-inspired insights into optimal sampling strategies, offers a framework for up-scaling measurements, and provides a down-scaling approach for integration of models and measurements. These techniques can be applied at different spatial and temporal scales to meet the needs of individual measurement campaigns. Thus, they are applicable to challenges likely to be encountered by many of those tasked with the design and coordination of monitoring networks throughout the Arctic.

Forrest M. Hoffman (forrest@climatemodeling.org), Jitendra Kumar (jkumar@climatemodeling.org), Stan D. Wullschleger (wullschlegsd@ornl.gov), Larry D. Hinzman (lhinzman@iarc.uaf.edu), Edward A. G. Schuur (tschuur@ufl.edu)

2 Introduction

The International Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) has documented strong evidence for warming of the Earth's climate over the last century and has attributed the increase in global temperatures primarily to rising anthropogenic greenhouse gas emissions (IPCC, 2007). Climate warming is projected to continue with globally important climate feedbacks and broad implications for sensitive ecosystems (Anisimov et al., 2007). Warming is projected to be especially pronounced at high latitudes and accompanied by significant regional impacts. Evidence of Arctic-wide responses is already being observed (Hinzman et al., 2005). Despite these potential implications, the Arctic has a limited record of low density observations. The Arctic Cimate Impact Assessment (ACIA) (2005) emphasized the need for studies of the complex and interacting processes of the atmosphere, sea ice, ocean, and terrestrial systems to improve the interpretation of past climate and projections of future climate. The Committee on Designing an Arctic Observing Network (2006) identified critical needs and gaps for observations in the Arctic. It recommended an Arctic Observing Network (AON) to satisfy current and future scientific needs and offered recommendations on key physical, biogeochemical, and human dimensions variables to monitor. Subsequently, The Arctic Observing Network Design and Implementation Task Force (2012) summarized these findings and explored options for a coordinated AON architecture. That report identified the following as overarching design strategy needs.

- Addressing observational requirements (accuracy, frequency, locations, etc.) with quantitative rigor, and
- identifying the architecture of a system-scale framework that will enable assessments of how particular observations would impact understanding and prediction issues or problems that span several components of the Arctic system.

While the report suggested three quantitative model-based assessment methods, primarily relying on repeated data-denial experiments employing dynamical data assimilation in complex process models to characterize the variability of individual indices or variables, it identified no general-purpose statistical modeling methodology that might usefully partition the spatial and temporal variance of an entire suite of environmental characteristics simultaneously and suggest representative sampling locations. Such a technique is described below.

Conducting systematic field observations and long-term monitoring of important processes are challenging, particularly in the Arctic. The value of such networks is strong as evidenced by findings made possible by the CALM, ITEX, and permafrost temperature (TSP) networks, to name just a few. Resource and logistical constraints limit the frequency and extent of observations, necessitating the development of a sampling strategy that objectively represents environmental variability at the desired spatial scale. Statistical design of the network, particularly the location of sampling sites, is critical for maximizing the representativeness of the sampled data, given a fixed number of sampling locations. Required is a methodology that provides a quantitative framework for stratifying sampling domains, informing site selection,

and determining the representativeness of measurements. This information is required for up-scaling and extrapolating point measurements to the larger landscape based on the similarity of its environmental characteristics.

Here we describe a quantitative, statistical approach to the design of spatially distributed monitoring networks. Given the availability of gridded datasets characterizing important environmental factors, sampling domains can be delineated and the representativeness of the monitoring network can be quantified in the context of a larger region or continent. This analysis can be carried out by varying the number of sampling sites in the network to quantitatively evaluate the incremental contribution of each site to the network and to understand the constituency of each potential site. We illustrate the approach for the design of a monitoring network for the State of Alaska using climate and permafrost-related characteristics for the present and a potential future that includes significant warming. We discuss network design in terms of site and network representativeness in the context of the large domain (the State of Alaska), and elaborate on extensions of the approach for understanding the constituency of individual sites and the use of ancillary logistics data for selecting affordable and accessible sites while minimizing compromises to representativeness. This method is independent of resolution, and may be applied across a handful of small sampling plots, within watersheds or basins, throughout a region, or across the entire pan-Arctic if adequate, continuous data are available for the domain in question.

3 Delineation of Quantitative Ecoregions

3.1 Ecoregion Concept

Ecoregions have been widely used to stratify geographic domains into nearly homogeneous land areas with respect to their geophysical, biological, and climatic characteristics. Since ecoregions are designed to correspond well with biome distributions and species ranges, they are frequently used as a framework for studying ecosystem structure and function. Qualitative and generalized ecoregion maps of the United States and the world have traditionally been developed by experts for studying ecosystem behavior or to define units for land management (Omernik, 1987; Olson and Dinerstein, 2002; Bailey and Hogg, 1986; Bailey, 2009). Hargrove and Hoffman (1999) used cluster analysis for quantitative delineation of ecoregions using a set of nine environmental characteristics for the conterminous United States and subsequently demonstrated its application for sampling network design, environmental niche modeling, and comparison of global model predictions (Hargrove and Hoffman, 2004; Hoffman et al., 2005). Nowacki and Brock (1995) and Gallant et al. (1995) produced ecoregion maps for the State of Alaska using two different expert-based methodologies, strongly focused on land form. Later, Nowacki et al. (2001) produced a "unified" ecoregion map—combining the two expert-based techniques—by considering limited data and in consultation with experienced ecologists, biologists, geologists, and regional experts. While useful for some purposes, such qualitative maps are based on the subjective expertise of the person or group developing them and suffer from various limitations (Hudson, 1992; Zhou, 1996).

The question of whether ecoregions can or should be developed using quantitative statistical methods or should rely upon human expertise has been a matter of debate among geographers (McMahon et al., 2001). In this study, the ecoregion concept is employed as a framework for environmental sampling, and a multivariate spatiotemporal cluster analysis technique is used to delineate these ecoregions.

3.2 Multivariate Spatiotemporal Clustering (MSTC)

The k-means algorithm (Hartigan, 1975) clusters a dataset of n observation vectors into a user-selected number of groupings or clusters (k). The algorithm begins by calculating the Euclidean distance of each observation to initial centroid vectors and classifies or assigns each observation to its nearest centroid. Each centroid vector is recalculated as the vector mean of all observations assigned to it. This classification and re-calculation process is iteratively repeated until fewer than some fixed proportion of observations change their cluster assignment between iterations. In the algorithm used here, convergence is assumed once fewer than 0.05% of the observations change cluster assignments. The results of the k-means algorithm are sensitive to the choice of initial centroids. Various heuristics may be employed for their selection, such as choosing initial centroids to have an even distribution within data space or to be spread along the edges of the distribution of observations. In this study, a multi-stage refinement method based on the work of Bradley and Fayyad (1998) is employed.

For geographic or spatial stratification applications, observation vectors consist of map cells, the dimensions of which are the biological or geophysical characteristics or variables under consideration. For spatiotemporal partitioning, observation vectors consist of map cells at different time periods. Hoffman and Hargrove (1999) developed a parallel version of the k-means algorithm for use on clusters of inexpensive personal computers (Hargrove et al., 2001), and this code was used in a meta-computing environment to cluster data using multiple supercomputers across the Internet (Mahinthakumar et al., 1999). Hoffman et al. (2008) later implemented improvements to accelerate convergence, handle empty cluster cases, and obtain initial centroids through a scalable implementation of the Bradley and Fayyad (1998) method. Kumar et al. (2011) extended this work to develop a fully distributed, highly scalable k-means parallel clustering tool for analysis of very large data sets, which was employed in this study.

3.3 Input Data Layers

The environmental characteristics that serve as input data layers are selected for inclusion in the analysis based on their expected predictive power for the environmental properties and processes of interest. The strength of that predictive power can be tested for individual or combinations of characteristics by performing a series of factorial analyses with and without them included. This analysis used a set of 37 environmental characteristics, or variables, shown in Table 1, from down-scaled general circulation model (GCM) results and observational data for the State of Alaska at a nominal resolution of 2 km $\times$ 2 km. These data were used to define a collec-

Table 1: The 37 characteristics or variables, averaged for 2000–2009 and 2090–2099, used in Multivariate Spatiotemporal Clustering (MSTC) for the State of Alaska.

Description	Number or Name	Units	Source
Monthly mean air temperature	12	°C	GCM
Monthly mean precipitation	12	mm	GCM
Day of freeze	mean	day of year	GCM
	standard deviation	days	
Day of thaw	mean	day of year	GCM
	standard deviation	days	
Length of growing season	mean	days	GCM
	standard deviation	days	
Maximum active layer thickness	1	m	GIPL
Warming effect of snow	1	°C	GIPL
Mean annual ground temperature at bottom of active layer	1	°C	GIPL
Mean annual ground surface temperature	1	°C	GIPL
Thermal offset	1	°C	GIPL
Limnicity	1	%	NHD
Elevation	1	m	SRTM

tion of ecoregions at multiple levels of division across two time periods for Alaska. Model results were averaged for the present (2000–2009) and the future (2090–2099). This analysis combined temperature, precipitation, and related bio-climatic projections from a five-model composite data set of down-scaled GCM results for the A1B emissions scenario Nakićenović et al. (2000) described by Walsh et al. (2008); corresponding snow and permafrost projections from the Geophysical Institute Permafrost Lab (GIPL) 1.3 permafrost dynamics model forced with the composite GCM results (Romanovsky and Marchenko, 2009); limnicity data based on the National Hydrography Dataset (NHD), pre-processed by Arp and Jones (2009); and elevation data from the Shuttle Radar Topography Mission (SRTM). The same limnicity and elevation data were used for both time periods. Because the units of measurement differ between variables, all data were standardized such that each variable had a mean of zero and a standard deviation of one prior to clustering to equalize the contribution from each predictor.

3.4 Alaska Ecoregions – A Case Study

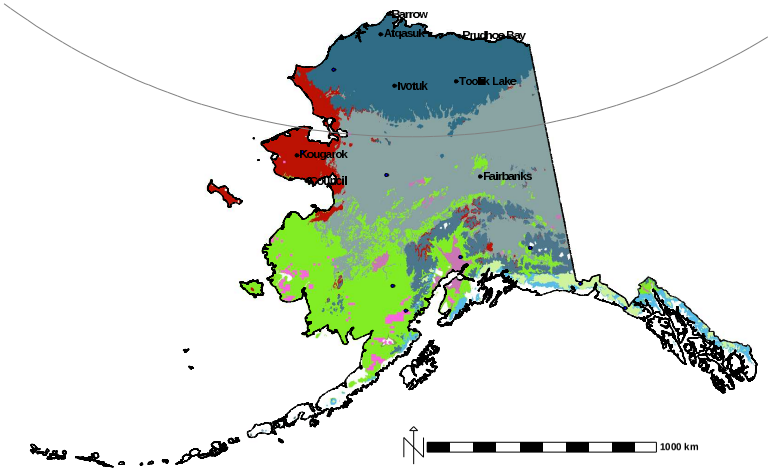
We applied the MSTC approach to derive ecoregions based on climate and topographic factors for the present and the future at multiple levels of division. The climate and topographic factors discussed above describe the environmental conditions of each map cell and are considered important drivers controlling vegetation and primary production. Thus, groupings or clusters of similarly characterized map cells delineated based on these variables define unique ecoregions. As demonstrated by Hargrove and Hoffman (2004), both present and projected future climate factors were included in the same analysis so that groups of similar cells were objectively

determined across space and through time. MSTC provides a basis for comparison of environmental conditions in the future with those in the present. Ecoregions constructed through this analysis may grow or shrink in spatial area and may shift across the landscape. At high levels of division or under extreme environmental change conditions, some present-day ecoregions may become extinct in the future (*i.e.*, shrink to zero spatial area), while others may exist only in the future (*i.e.*, have no analog in the present). This quantitative delineation of ecoregions across space and through time facilitates assessment of the magnitude of change between present and future environmental conditions and enables the evaluation of the ecological implications of climate change scenarios. From a conservation perspective, this methodology maps changing habitats and species at risk from climate change (Saxon et al., 2005). From a field sampling perspective, this methodology identifies regions fostering potentially vulnerable ecosystems or supporting large and vulnerable carbon stores that may be sensitive to climate change (McGuire et al., 2009; Chapin et al., 2010). Such ecoregions warrant intense observation and benefit from careful, quantifiable, and defensible sampling network design strategies.

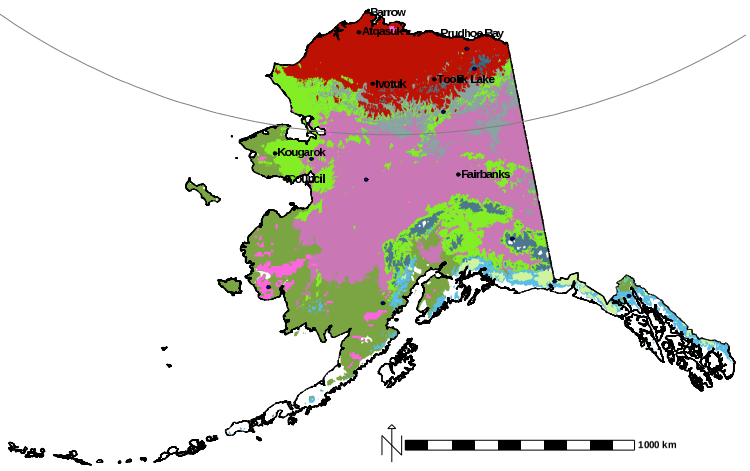
Expert-derived ecoregion maps are static and have boundaries based on subjective consideration of geographic properties and expert judgment. In contrast, statistically derived ecoregions can vary with time and are delineated in the data space or state space representing all the characteristics under consideration. Moreover, the state space resolution can be varied by selecting different numbers of clusters. Figures 1(a) and 1(b) contain maps of the 10 quantitatively defined, most-different Alaskan ecoregions for the present and future, respectively. These quantitative ecoregions correspond well with commonly identifiable ecosystem types, like tundra, taiga, and rainforests (Hoffman et al., submitted). The cluster centroid of each ecoregion represents the mean value of all the characteristics or state variables for that ecoregion. Increasing the selected number of clusters in the k-means algorithm allows the definition of a larger number of more specifically defined, less generalized ecoregions. For example, Figures 1(c) and 1(d) contain maps of the 20 quantitatively defined, most-different Alaskan ecoregions for the present and future, respectively. By continuing to increase the level of division, the state space resolution can be further increased. In some cases, a natural hierarchy among ecosystems in the landscape is manifested as the level of division is increased (Hoffman et al., submitted).

4 Site Selection

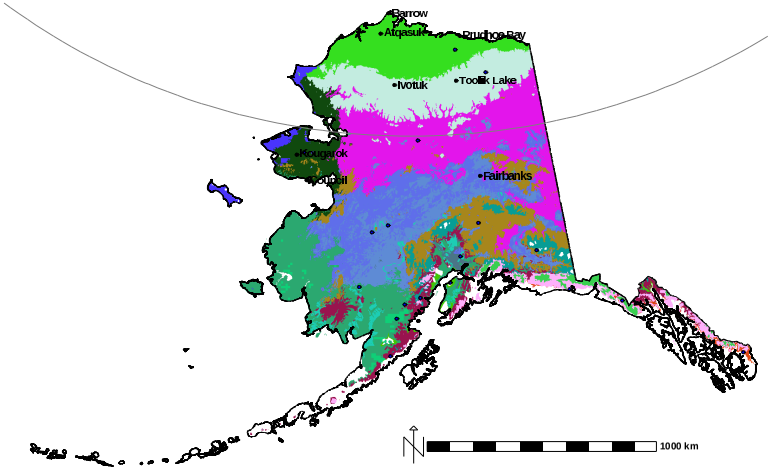
Selection of sampling locations for long term monitoring of ecosystem properties and processes should be guided by an objective, quantitative, systematic, and defensible methodology. Instead, sampling locations in large-scale networks have often been established in opportunistic, political, or logistically-driven ways only, resulting in unquantified representation of heterogeneity, biased sampling, uncharacterized uncertainty, and undirected network growth. Finite resources and logistical constraints limit the spatiotemporal frequency and extent of environmental observations, necessitating the development of a systematic sampling strategy to objectively represent environmental variability at the desired spatial scale. An appropriately designed ob-



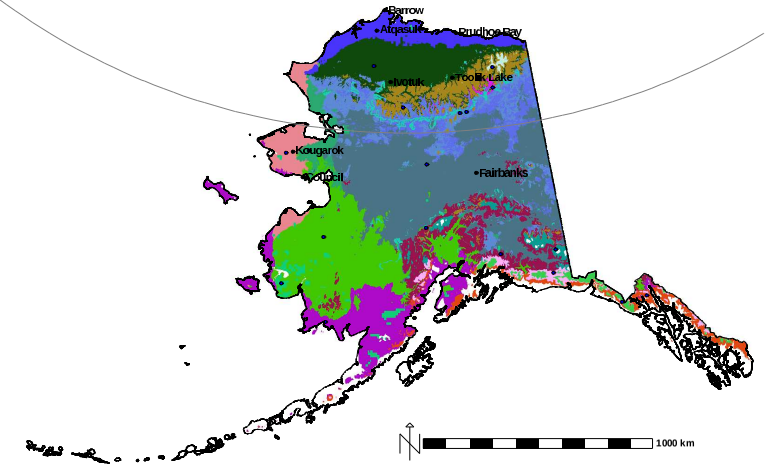
(a) 10 ecoregions, present (2000–2009)



(b) 10 ecoregions, future (2090–2099)



(c) 20 ecoregions, present (2000–2009)



(d) 20 ecoregions, future (2090–2099)

Figure 1: The 10 (a and b) and 20 (c and d) most-different quantitatively defined ecoregions for the State of Alaska in the present (a and c) and future (b and d) decades were derived from 37 variables and are shown using random colors. Realized centroids, map locations most closely approximating the mean value within an ecoregion of all the 37 variables, are indicated by the blue dot in each ecoregion.

servation strategy should be employed to quantitatively delineate sampling domains, sites, and frequencies. The National Science Foundation's (NSF's) National Ecological Observatory Network (NEON) adopted the objective, data-based methodology described above to define 20 optimal sampling domains across the conterminous United States (Keller et al., 2008; Schimel et al., 2007). Accurate characterization of the landscape and translation of data collected in the field and laboratory into useful datasets, process algorithms, and model parameters requires classification of the landscape into discrete units based on ecological, hydrological, and geological properties. In much the same way that ecologists develop ecoregions, geologists often classify landscape areas into geomorphological units based on their geophysical and hydrological features. For complex and evolving landscapes featuring interacting vegetation and geomorphological dynamics responding to changes in climate, such as in the Arctic, these stratification concepts may be unified to produce *biogeomorphic units* at relevant spatial scales for landscape characterization, identification of ecological and geomorphological processes, assessing the representativeness of measurements, and providing a framework for scaling measurements and model parameters to larger domains.

An important aspect of site selection and the up- and down-scaling approach to integration of models, observations, and process studies is the estimation of *representativeness*. The MSTC methodology described above for landscape characterization offers useful metrics for indicating the representativeness of sites, measurements, and model parameters. Hargrove et al. (2003) described this technique for understanding the representativeness of a sampling network based on a suite of environmental gradients considered to be useful proxies for the characteristics being measured. They applied the technique to quantify the representativeness of AmeriFlux Network. Maps identifying poorly represented regions can be produced, suggesting where new measurements should be taken to maximize the value of limited observations in a sparse sampling network. As discussed earlier, since the cluster centroid represents the mean value of all the state variables in an ecoregion, the realized centroid for an ecoregion is the location that best represents the combination of environmental conditions of the entire ecoregion. Therefore, statistically defined realized centroids, indicated by blue dots in each ecoregion in Figure 1, are the optimal sampling locations for each ecoregion. Logistical constraints—including accessibility, availability of electric power and telecommunications infrastructure, and geologic stability—may prevent establishment of sampling sites at such optimal locations, particularly in an Arctic environment. Nevertheless, the MSTC Ecoregion framework can be used to strike the best compromise between logistical practicality and domain representativeness by performing a post-hoc analysis using logistical constraints, like distance to roads or power, difficulty of accessibility for construction and maintenance, availability of power and communications, etc. This approach provides a means for quantifying the representativeness of measurements taken at sub-optimal locations, either within an ecoregion or across any larger domain for which the desired state variables are available.

5 Quantifying Representativeness

While most *in situ* field measurements are made at relatively small, individual geographic points, ecosystem processes operate at many scales. In order to utilize limited point measurements at larger spatial and temporal scales for input to or evaluation of process modeling or for estimating landscape-scale characteristics, the representativeness of those measurements must be quantified in the context of a heterogeneous and evolving landscape. A useful representativeness metric is one that can inform the selection of sampling locations, up-scaling of point measurements, down-scaling of remote sensing data, and extrapolation of measurements to unsampled domains. The representativeness metric described by Hargrove et al. (2003) provides a unit-less, relative measure of the dissimilarity between the ecoregion of interest, which may contain a sampling site, and any other ecoregion. It is calculated as the Euclidean distance between two ecoregion centroids within the standardized n-dimensional state space. Ecoregions with similar combinations of environmental conditions will have centroids located near to each other in state space. Therefore, the Euclidean distance between those centroids will be small, representing a low dissimilarity or high representativeness measure. Meanwhile, ecoregions with very different combinations of environmental conditions will have centroids located far from each other in state space, resulting in a large Euclidean distance between them. Such ecoregions will have a high dissimilarity or low representativeness measure. To best capture the detailed heterogeneity at the scale of interest, this ecoregion-based representativeness should be calculated using MSTC Ecoregions with a large number of divisions (*i.e.*, a large value of k).

While Hargrove et al. (2003) calculated representativeness in the context of ecoregions; however, this same approach can be applied to every map cell projected individually onto the n-dimensional state space used to perform the cluster analysis that produced MSTC Ecoregions. This *point-based representativeness* metric captures the full range of heterogeneity in the combinations of environmental conditions, providing a continuously varying measure of dissimilarity for every map cell with respect to a map cell of interest, which may contain a sampling location. When a single ecoregion centroid or map cell of interest is considered, a map of *site representativeness* can be produced. However, multiple ecoregions or map cells of interest may be considered simultaneously, for instance, to provide a quantitative measure of the representativeness of an array or network of sampling sites. The result is a map of *network representativeness* for which the dissimilarity measure for every ecoregion centroid or map cell is the Euclidean distance between that point and the nearest ecoregion centroid or map cell of interest (*i.e.*, the minimum value from a stack of site representativeness maps, one for each ecoregion centroid or map cell containing a measurement site). This representativeness metric, whether ecoregion- or point-based, can be calculated not only between different geographic points in space, but also between different (or the same) geographic points through time. For example, the Euclidean distance between the present combination of environmental conditions and those of the future for any single map cell represents a measure of the magnitude of environmental change over time. Therefore, with this metric it is possible to calculate not only the present-day representativeness of measurements from

a site, but also the future representativeness of those present-day measurements, based on future projections of the state variables used in the analysis.

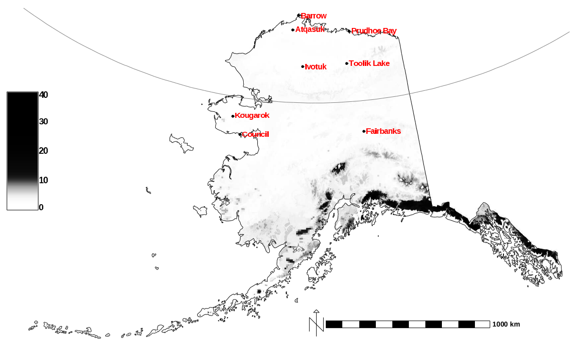
6 Network Representativeness

A monitoring network often consists of a geographically distributed constellation of measurement sites or may be locations where samples are collected for further analysis in the laboratory. Quantifying the representativeness of the network as a whole is important for optimal network design, to avoid unnecessary duplication and to maximize the coverage of the monitoring network. By combining multiple maps of site representativeness for every sampling location, and calculating the minimum value for every map cell, maps of network representativeness are produced. Figures 2(a) and 2(b) contain maps of ecoregion-based network representativeness for all eight candidate sampling sites for the present and future time periods, respectively. Similarly, Figures 3(a) and 3(b) contain maps of point-based network representativeness for the same eight candidate sampling sites for the present and future time periods, respectively. White to light gray land areas are well-represented by the network of sites, while dark gray to black land areas are poorly represented by the network of sites. If the objective were to maximize the coverage of all environments in the State of Alaska, the next sampling location should be chosen within the darkest land areas shown in the map.

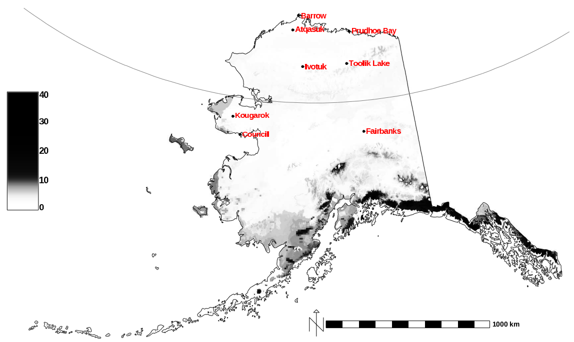
7 Conclusion

Systematic sampling strategies are essential for understanding ecosystem responses to climate change and informing model development. In the harsh Arctic environment—where climate change appears to be most rapidly affecting sensitive ecosystems and vulnerable, carbon-rich permafrost—filling critical gaps in observations is expensive and technically challenging. To fully explore the regional and global implications of climate change in the Arctic, global Earth System Models must capture the important processes and feedbacks. Such models must be developed based on a rich body of relevant observational data as representative as possible of multiple spatial and temporal scales. Meanwhile, finite resources and logistical constraints place restrictions on the number of sampling sites, spatial extent, frequency, and types of measurements that can be collected. This study proposes a quantitative, data-based methodology for stratifying sampling domains, informing site selection, and determining the representativeness of measurement sites and sampling networks.

Multivariate spatiotemporal clustering (MSTC), based on k-means cluster analysis, was applied to down-scaled general circulation model (GCM) results and observational data for the State of Alaska at a nominal resolution of 2 km $\times$ 2 km to define a set of ecoregions at multiple levels of division across two decadal time periods. Maps of ecoregions for the present (2000–2009) and future (2090–2099) were produced, showing how combinations of 37 environmental conditions are distributed across Alaska and how these combinations shift as a result of projected

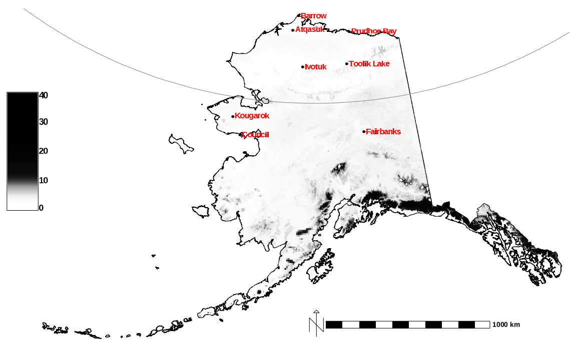


(a) Ecoregion-based network representativeness of eight sites for the present period

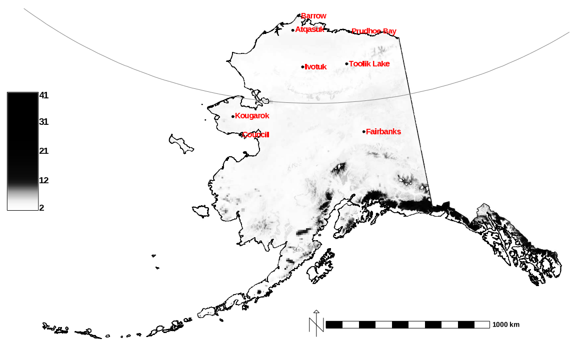


(b) Ecoregion-based network representativeness of eight sites for the future period

Figure 2: Representativeness maps for a network of eight sites for the present and future time periods. White to light gray land areas are well-represented by the network of sites, while dark gray to black land areas are poorly represented by the network of sites.



(a) Point-based network representativeness of eight sites for the present period



(b) Point-based network representativeness of eight sites for the future period

Figure 3: Representativeness maps for a network of eight sites for the present and future time periods. White to light gray land areas are well-represented by the network of sites, while dark gray to black land areas are poorly represented by the network of sites.

climate change in the 21st century. Using this statistical approach, optimal sampling locations, called realized centroids, were identified for each ecoregion at every level of division. In addition, the resulting geographic shifts and changes in areal distribution of ecoregions suggested that some environments may disappear, many will be redistributed, and new ones will appear in the coming century. This analysis provides insights into the identification of the most sensitive and potentially vulnerable Arctic ecosystems and suggests optimal monitoring network strategies for observing those changes. The Euclidean distance within the 37-dimensional state space used for MSTC provides a metric for representativeness. Gray-scale maps of representativeness, showing the similarity of every map cell to a list of eight candidate samples locations near town sites in Alaska, were produced for each site. Taken together, these analysis products provide model-inspired insights into optimal sampling strategies across space and through time, and these same techniques can be applied at different spatial and temporal scales to meet the needs of individual measurement or monitoring campaigns.

The representativeness of a sampling network is best maximized before the network is deployed. Even if additional "optimized" sites are added to an existing network, it will require many more additions to approach the theoretical maximum representativeness for a given number of initial sites. It is difficult, with only the sequential addition of new optimized sites, to achieve the same representativeness once some sampling sites have been established. Representativeness resulting from such network "repairs" rarely ever equal the representativeness of a network initially designed *de novo* with that same number of sampling sites. Even if the network is to be constructed in stages, it is best to design site placement using the final, ultimate complement of sites and to operate sub-optimally until the full network can be completed. Otherwise, many more sites will have to be added to the existing network in order to achieve the same representativeness than could otherwise have been designed in initially.

Cluster analysis and n-dimensional data space regressions offer quantitative methods for up-scaling and extrapolating measurements to land areas within and beyond the sampling domain and provide a down-scaling approach to the integration of models, observations, and process studies. The accuracy of the up-scaled data will be higher for areas represented well by the monitoring network and lower for areas that are poorly represented. At a large scale, these techniques are useful for delineating distinct, broad regions and optimal measurement sites. However, this methodology can also be applied at finer spatiotemporal scales, with inclusion of other geophysical characteristics and remote sensing data, to inform measurement frequency and site selection within these broader ecoregions.

Acknowledgments

This research was sponsored by the Climate and Environmental Sciences Division (CESD) of the Office of Biological and Environmental Research (BER) within the U.S. Department of Energy (DOE) Office of Science. This research used resources of the Center for Computational Sciences at Oak Ridge National Laboratory, which

is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

References

IPCC. Summary for Policymakers. In Susan Solomon, Dahe Qin, Martin Manning, Zhenlin Chen, Melinda Marquis, Kristen B. Averyt, Melinda Tignor, and Henry L. Miller, editors, *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge, United Kingdom and New York, NY, USA, 2007. Cambridge University Press. ISBN 978-0-521-88009-1 hardback; 978-0-521-70596-7 paperback.

O. A. Anisimov, D. G. Vaughan, T. V. Callaghan, C. Furgal, H. Marchant, T. D. Prowse, H. Vilhjálmsson, and J. E. Walsh. Polar regions (Arctic and Antarctic). In M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, editors, *Climate Change 2007: Impacts, Adaptation and Vulnerability*, pages 653–685. Cambridge University Press, Cambridge, 2007.

Larry Hinzman, Neil Bettez, W. Bolton, F. Chapin, Mark Dyurgerov, Chris Fastie, Brad Griffith, Robert Hollister, Allen Hope, Henry Huntington, Anne Jensen, Gensuo Jia, Torre Jorgenson, Douglas Kane, David Klein, Gary Kofinas, Amanda Lynch, Andrea Lloyd, A. McGuire, Frederick Nelson, Walter Oechel, Thomas Osterkamp, Charles Racine, Vladimir Romanovsky, Robert Stone, Douglas Stow, Matthew Sturm, Craig Tweedie, George Vourlitis, Marilyn Walker, Donald Walker, Patrick Webber, Jeffrey Welker, Kevin Winker, and Kenji Yoshikawa. Evidence and implications of recent climate change in Northern Alaska and other Arctic regions. *Clim. Change*, 72(3):251–298, 2005. ISSN 0165-0009. doi:10.1007/s10584-005-5352-2. URL `http://dx.doi.org/10.1007/s10584-005-5352-2`.

The Arctic Cimate Impact Assessment (ACIA). *Arctic Climate Impact Assessment*. Cambridge University Press, 2005. ISBN 9780521865098. URL `http://www.acia.uaf.edu/pages/scientific.html`.

National Research Council Committee on Designing an Arctic Observing Network. *Towards an Integrated Arctic Observing Network*. The National Academies Press, 2006. ISBN 9780309100526. URL `http://www.nap.edu/openbook.php?record_id=11607`.

The Arctic Observing Network Design and Implementation Task Force. Designing, optimizing, and implementing an Arctic Observing Network. Technical report, Study of Environmental Arctic Change (SEARCH), Fairbanks, Alaska, 2012.

James M. Omernik. Ecoregion of the conterminuous United States. *An. Assoc. Amer. Geog.*, 77(1):118–125, 1987. doi:10.1111/j.1467-8306.1987.tb00149.x.

David M. Olson and Eric Dinerstein. The global 200: Priority ecoregions for global conservation. *Annals of the Missouri Botanical Garden*, 89(2):199–224, April 2002. ISSN 00266493. URL `http://www.jstor.org/stable/3298564`.

Robert G. Bailey and Howard C. Hogg. A world ecoregions map for resource reporting. *Environ. Conserv.*, 13(3):195–202, September 1986. doi: 10.1017/S0376892900036237.

Robert G. Bailey. Ecoregions of the United States. In *Ecosystem Geography*, Statistics for Social and Behavioral Sciences, pages 93–114. Springer New York, 2009. ISBN 978-0-387-89516-1. doi:10.1007/978-0-387-89516-1_7. URL `http://dx.doi.org/10.1007/978-0-387-89516-1_7`.

William W. Hargrove and Forrest M. Hoffman. Using multivariate clustering to characterize ecoregion borders. *Comput. Sci. Eng.*, 1(4):18–25, July 1999. doi: 10.1109/5992.774837.

William W. Hargrove and Forrest M. Hoffman. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environ. Manage.*, 34 (Supplement 1):S39–S60, April 2004. doi:10.1007/s00267-003-1084-0.

Forrest M. Hoffman, William W. Hargrove, David J. Erickson, and Robert J. Oglesby. Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models. *Earth Interact.*, 9(10):1–27, August 2005. doi:10.1175/EI110.1.

Gregory Nowacki and Terry Brock. Ecoregions and Subregions of Alaska, EcoMap Version 2.0. map, USDA Forest Service, Alaska Region, Juneau, AK, 1995. URL `http://www.fs.fed.us/land/ecosysmgmt/index.html`. Scale 1:5,000,000.

A. L. Gallant, E. F. Binnian, J. M. Omernik, and M. B. Shasby. Ecoregions of Alaska. Professional paper 1567, U.S. Geological Survey, 1995. URL `http://agdcftp1.wr.usgs.gov/pub/projects/fhm/ecoregmeta.html`.

Gregory Nowacki, Page Spencer, Michael Fleming, Terry Brock, and Torre Jorgenson. Ecoregions of Alaska: 2001. Open-file report 02-297 (map), U.S. Geological Survey, 2001. URL `http://agdcftp1.wr.usgs.gov/pub/projects/fhm/akecoregions.htm`.

Berman D. Hudson. The soil survey as paradigm-based science. *Soil Sci. Soc. Am. J.*, 56(3):836–841, 1992. doi:10.2136/sssaj1992.03615995005600030027x. URL `https://www.soils.org/publications/sssaj/abstracts/56/3/836`.

Yingchun Zhou. An ecological regionalization model based on NOAA/AVHRR data. *International Archives of Photogrammetry and Remote Sensing*, XXXI, Part B4: 1001–1006, 1996.

Gerard McMahon, Steven M. Gregonis, Sharan W. Waltman, James M. Omernik, Thor D. Thorson, Jerry A. Freeouf, Andrew H. Rorick, and James E. Keys. Developing a spatial framework of common ecological regions for the conterminous united states. *Environ. Manage.*, 28(3):293–316, 2001. ISSN 0364-152X. doi: 10.1007/s0026702429. URL `http://dx.doi.org/10.1007/s0026702429`.

J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.

Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 91–99, San Francisco, CA, USA, July 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

Forrest M. Hoffman and William W. Hargrove. Multivariate geographic clustering using a Beowulf-style parallel computer. In Hamid R. Arabnia, editor, *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA '99)*, volume III, pages 1292–1298. CSREA Press, June 1999. ISBN 1-892512-11-4.

William W. Hargrove, Forrest M. Hoffman, and Thomas Sterling. The do-it-yourself supercomputer. *Sci. Am.*, 265(2):72–79, August 2001. URL `http://www.sciam.com/article.cfm?articleID=000E238B-33EC-1C6F-84A9809EC588EF21`.

Gnanamanika Mahinthakumar, Forrest M. Hoffman, William W. Hargrove, and Nicolas T. Karonis. Multivariate geographic clustering in a metacomputing environment using Globus. In *Supercomputing '99: Proceedings of the 1999 ACM/IEEE conference on Supercomputing (CDROM)*, Supercomputing '99, New York, NY, USA, November 1999. ACM Press. ISBN 1-58113-091-0. doi: 10.1145/331532.331537.

Forrest M. Hoffman, William W. Hargrove, Richard T. Mills, Salil Mahajan, David J. Erickson, and Robert J. Oglesby. Multivariate Spatio-Temporal Clustering (MSTC) as a data mining tool for environmental applications. In Miquel Sànchez-Marrè, Javier Béjar, Joaquim Comas, Andrea E. Rizzoli, and Giorgio Guariso, editors, *Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Society (iEMSs 2008)*, pages 1774–1781, July 2008. ISBN 978-84-7653-074-0.

Jitendra Kumar, Richard Tran Mills, Forrest M. Hoffman, and William W. Hargrove. Parallel k-means clustering for quantitative ecoregion delineation using large data sets. In Mitsuhisa Sato, Satoshi Matsuoka, Peter M. Sloot, G. Dick van Albada, and Jack Dongarra, editors, *Proceedings of the International Conference on Computational Science (ICCS 2011)*, volume 4 of *Procedia Comput. Sci.*, pages 1602–1611. Elsevier, Amsterdam, June 2011. doi:10.1016/j.procs.2011.04.173. URL `http://www.sciencedirect.com/science/article/B9865-52VR1G5-7X/2/898a65768000cb4b8efa46e647a60ae6`.

Nebojša Nakićenović, Joseph Alcamo, Gerald Davis, Bert de Vries, Joergen Fenhann, Stuart Gaffin, Kenneth Gregory, Arnulf Grübler, Tae Yong Jung, Tom Kram,

Emilio Lebre La Rovere, Laurie Michaelis, Shunsuke Mori, Tsuneyuki Morita, William Pepper, Hugh Pitcher, Lynn Price, Keywan Riahi, Alexander Roehrl, Hans-Holger Rogner, Alexei Sankovski, Michael Schlesinger, Priyadarshi Shukla, Steven Smith, Robert Swart, Sascha van Rooijen, Nadejda Victor, and Zhou Dadi. Special report on emissions scenarios. In Nebojša Nakićenović and Robert Swart, editors, *A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*, page 570. Cambridge University Press, Cambridge, United Kingdom, July 2000. ISBN 92-9169-113-5.

John E. Walsh, William L. Chapman, Vladimir Romanovsky, Jens H. Christensen, and Martin Stendel. Global climate model performance over Alaska and Greenland. *J. Clim.*, 21(23):6156–6174, December 2008. doi:10.1175/2008JCLI2163.1.

Vladimir E. Romanovsky and Sergei Marchenko. The GIPL permafrost dynamics model. Technical report, University of Alaska, Fairbanks, Alaska, May 2009. URL `http://www.snap.uaf.edu/files/TheGIPL-1Model-final.pdf`.

Christopher D. Arp and Benjamin M. Jones. Geography of Alaska lake districts: Identification, description, and analysis of lake-rich regions of a diverse and dynamic state. Scientific Investigations Report 2008-5215, U.S. Geological Survey, 4210 University Dr., Anchorage, Alaska 99508, January 2009. URL `http://pubs.usgs.gov/sir/2008/5215/`.

Earl Saxon, Barry Baker, William Hargrove, Forrest Hoffman, and Chris Zganjar. Mapping environments at risk under different global climate change scenarios. *Ecol. Lett.*, 8:53–60, 2005. doi:10.1111/j.1461-0248.2004.00694.

A. David McGuire, Leif G. Anderson, Torben R. Christensen, Scott Dallimore, Laodong Guo, Daniel J. Hayes, Martin Heimann, Thomas D. Lorenson, Robie W. Macdonald, and Nigel Roulet. Sensitivity of the carbon cycle in the Arctic to climate change. *Ecol. Monogr.*, 79(4):523–553, November 2009. doi:10.1890/08-2025.1.

F. S. Chapin, A. D. McGuire, R. W. Ruess, T. N. Hollingsworth, M. C. Mack, J. F. Johnstone, E. S. Kasischke, E. S. Euskirchen, J. B. Jones, M. T. Jorgenson, K. Kielland, G. P. Kofinas, M. R. Turetsky, J. Yarie, A. H. Lloyd, and D. L. Taylor. Resilience of Alaska's boreal forest to climatic change. *Can. J. Forest Res.*, 40(7):1360–1370, July 2010. doi:10.1139/X10-074.

Forrest M. Hoffman, Jitendra Kumar, William W. Hargrove, and Richard T. Mills. Representativeness-based sampling network design for the Arctic. *Landscape Ecol.*, submitted.

Michael Keller, David Schimel, William Hargrove, and Forrest Hoffman. A continental strategy for the National Ecological Observatory Network. *Front. Ecol. Environ.*, 6(5):282–284, June 2008. doi:10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2. Special Issue on Continental-Scale Ecology.

David Schimel, William Hargrove, Forrest Hoffman, and James McMahon. NEON: A hierarchically designed national ecological network. *Front. Ecol. Environ.*, 5(2): 59, March 2007. doi:10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2.

William W. Hargrove, Forrest M. Hoffman, and Beverly E. Law. New analysis reveals representativeness of the AmeriFlux Network. *Eos Trans. AGU*, 84(48): 529, 535, December 2003. doi:10.1029/2003EO480001.