Generating data ensembles over a model grid from sparse climate point measurements

Meredith Franklin\textsuperscript{1,2}, V. Rao Kotamarthi\textsuperscript{1}, Michael I. Stein\textsuperscript{2} and David R Cook\textsuperscript{1}

\textsuperscript{1}Argonne National Laboratory, Argonne, IL 60439, USA
\textsuperscript{2}University of Chicago Chicago, IL 60637, USA

E-mail: vrkotamarthi@anl.gov

Abstract. Parameterizations in high-end climate models can be evaluated with observed climate and meteorological data. Often this evaluation is achieved by averaging measured quantities over time and space to match the spatial and temporal resolution of a gridded climate model without any regard for the statistical errors in doing so. We present a statistical method to interpolate sparsely located surface measurements into a uniform spatial grid representative of global and regional models of the atmosphere and climate. This method provides estimates of mean values over the entire domain containing the measurement sites as well as the uncertainty in the estimated quantities at the grid locations through multiple simulations to create “data ensembles.” We demonstrate this method using measurements of surface sensible heat flux over the southern Great Plains region gathered through the Department of Energy’s Atmospheric Radiation Measurement (ARM) program. Application to the numerous other climate variables measured through the ARM program and the development of a software tool to streamline data management and implementation of the statistical models is also discussed.

1. Introduction

Continued improvement in the performance of climate models requires evaluation of their various parameterizations, particularly for processes driven by subgrid features such as orography, soil characteristics, vegetation, and land use. This can be achieved through validation with observed climate-related process measurements over decadal-length scales \cite{1}. Such observations enable validation of models over large spatial scales (regional to continental) and large temporal scales (seasonal).

The Atmospheric Radiation Measurement Program Climate Research Facility (ARM ACRF) hosts a suite of instruments that continually collect data on such measurements at high temporal frequency \cite{2}. While the ARM sites are spaced throughout a region in the U.S. southern Great Plains (SGP) designed to simulate the grid size of a climate model, their spatial scale is not suitable for model evaluation. Sites are widely scattered, with a heavily instrumented central location (figure 1) and ancillary sites arranged to cover different land use and topographic features in the region. The data collected ranges from surface measurements to remote sensing measurements used to generate vertical (3-D) distributions of parameters such as wind profiles.

Measurements are taken on different temporal scales and exhibit varying degrees of spatial heterogeneity. A given quantity, such as surface sensible heat flux, can vary widely across the domain from location to location (figure 2). Much of the variability is due to the variations in topography, local soil conditions, and geographic features. Variables such as temperature may show less deviation over a diurnal cycle at the different SGP site; however, cloud cover, soil moisture, and precipitation show similarly large spatial and temporal deviations.

In an initial sensitivity study, we examined the significance of using ARM data to constrain the Single Column Atmosphere Model (SCAM), which represents a single column of
the full Community Climate System Model (CCSM) distributed with the latest version of the full National Center for Atmospheric Research Community Atmosphere Model (CAM3).

**Figure 1.** The ACRF SGP site. Around the central facility near Lamont, OK, measurements are made at extended facilities and a few intermediate facilities.

**Figure 2.** Diurnal variability of surface sensible heat flux recorded at 13 stations within the ARM ACRF SGP site.

The difference in the surface latent heat flux under two identical simulations is shown in figure 3, with the surface latent heat fluxes set to the derived surface latent flux (figure 3a) or constrained with measurements (figure 3b). During daytime hours, the latent heat flux in the unconstrained model is nearly 50% higher than that measured during the ARM 1995 intensive operational period (IOP).

**Figure 3.** (a) SCAM unconstrained and (b) SCAM constrained with latent heat flux.

The difference in latent heat flux between the constrained and unconstrained model can be attributed to the SCAM’s parameterization of planetary boundary layer, which may account for nearly 50% water vapor under different initializations. Thus, a strong constraint on surface heat flux would be a valuable tool for evaluating the model-calculated moisture distribution.

While the 14 ARM Energy Balance Bowen Ratio (EBBR) sites in the SGP make measurements of surface heat flux, to evaluate climate model-calculated quantities, data from the sites must be converted from a point-process to a uniform grid surface by spatio-temporal interpolation. Furthermore, it is important to have a realistic approximation of uncertainty in the interpolated grid surface.

In recent years there has been intensive research in developing statistical methods for spatio-temporal interpolation [3-5] of environmental monitoring data. These methods are based on the construction of covariance matrices that describe an observed space-time process, applied
to predict values at unobserved locations and times. Often only a single prediction is generated at any unobserved location, but as Schneider [6] notes, it is extremely important to have multiple predictions in order to quantify uncertainty in the estimates. Furthermore, it is useful to evaluate climate model parameterizations with ensemble runs [7]; this can be facilitated by using “ensembles” of observed data.

We present a statistical approach for spatio-temporal interpolation of ARM data to a uniform spatial grid over the SGP domain. We illustrate our approach with sensible heat flux (h) measurements obtained from the 14 EBBR sites. Finally, by generating multiple realizations of the interpolated grid surface we quantify the uncertainty in our predictions and create data ensembles for evaluating climate model parameterizations.

2. Methods and results

All ARM data require varying degrees of preprocessing before any modeling can be conducted. In certain cases, quality flags have been included in the data product, while in others careful inspection and understanding of the instruments collecting the data are necessary to remove anomalies.

2.1 Data processing and gap-filling of flux data

The 14 SGP EBBR sites (figure 4) produce 30-minute estimates of vertical fluxes of surface sensible and latent heat. The flux estimates are calculated by the Bowen ratio technique, using measured net radiation, soil surface heat flux, and the vertical gradients of temperature and relative humidity. Several known issues with sensible and latent heat flux using the EBBR system were addressed before modeling these data. First, if the measured Bowen ratio, from which h and lambda e are derived, is close to -1, the fluxes become very large:

\[
\lambda e = -\frac{(Q^* + G)}{1 + B} \quad \text{and} \quad h = B\lambda e,
\]

where \(Q^*\) is net radiation, \(G\) is average ground heat transfer and \(B\) is the Bowen ratio. Second, because of the nature of the EBBR system, the data are reliable only when wind is coming over land that has sufficient fetch. Depending on the land type surrounding the instrument, there may be many instances where the flux data are questionable. A comparison of the data over “good” and “poor” fetch indicated a statistically significant difference in the flux values depending on the wind direction. Moreover, instrument issues can cause short gaps in the data series. After removing data that had any of the above data quality issues, gap filling was necessary to obtain a complete time series for modeling.

As shown in figure 2, sensible heat flux is highly variable by site, and thus we approached filling in missing values on a site-by-site basis. Heat flux is related to net radiation by \(Q^* = h + \lambda e + G\), where during daytime \(h \perp Q^*\) will be the dominant relationship for sensible heat flux. Thus we use the measured \(Q^*\) to fill gaps in daytime sensible heat flux. To fill in values for which both \(h\) and \(Q^*\) were missing during the daytime, we empirically calculated the solar radiation received at each EBBR site (insolation) by \(R = S_0 \phi \cos \theta\), where \(S_0\) is the solar constant, \(\phi\) is the corrected Earth-sun distance, and \(\cos \theta\) is the cosine of the zenith angle. We then regressed \(Q^*\) on \(R\) and used this relationship to calculate \(Q^*_2\). Missing \(h\) for which measured \(Q^*\) was also missing were filled with \(Q^*_2\). To predict the remaining missing \(h\) (occurring at night), we regressed available \(h\) on \(Q^*\), adjusting for diurnal and seasonal patterns with cubic regression spline functions and applied this fitted model to scale \(Q^*\) and \(Q^*_2\) to be more
representative of nighttime h. A 7-day period at site E2 where missing sensible heat flux measurements were gap-filled with this method is shown in figure 5.

2.2 Spatio-temporal interpolation: from point process to model grid

We characterize the space-time process of sensible heat flux over the SGP domain, D, by a nonseparable covariance function:

\[
\text{cov}\{Z(s_1,t_1), Z(s_2,t_2)\} = K(s_1 - s_2, t_1 - t_2)
\]

for N space-time coordinates \((s_1, t_1), \ldots (s_N, t_N)\) for \(s \in D, t \geq 0\) where

\[
K(s,t) = \int_{-\pi}^{\pi} S_0(\omega)e^{i\omega t} d\omega + \int_{-\pi}^{\pi} S(\omega)C(\{s\}^{\gamma(\omega)})e^{i\omega s}\theta(\omega) + i\omega t d\omega
\]

is spectral in time and has an isotropic covariance function [5, 8].

To fit this model it is necessary to have a stationary Gaussian process. For the sensible heat flux data, this was achieved by removing 15 diurnal and 12 seasonal (monthly) frequencies for each year of data from each site. The time series was then converted to the spectral domain by a discrete fast Fourier transform (DFFT). Other parameters in the model, including the coherence between spectra at the sites \(\gamma(\omega)\) in \(C(\{s\}^{\gamma(\omega)})\), the phase \(\theta(\omega)\) in \(e^{i\omega s}\theta(\omega)\) and \(S(\omega)\) were defined with cubic spline basis functions. The isotropic spatial covariance function, C was chosen to be a Matérn function with a smoothness parameter of 1.5 (see [5]). The coordinates of each site were transformed to the UTM zone 14N for ease in computing distances between sites. A grid consisting of 1x1 km grid cells over the entire SGP domain was constructed to serve as the interpolation surface of the ARM data. This fine spatial grid fits within four 2.5-degree Community Climate System Model (CCSM) grid cells.

With all of the parameters initialized, the combined model \(K(s,t)\) was fit through optimization to maximize its likelihood. This space-time process was used to predict h at each unobserved grid cell in the 1 km grid (figure 6).
2.3 Data ensembles
We estimated uncertainty in our predictions by computing multiple simulations [8] under the modeling framework described above. These simulations are conditional on the ARM data we observe and the maximum likelihood estimates (MLE) of the spatio-temporal covariance function.

Essentially, the process $Z(s^*, t^*)$, where $s^*$ is the $(x, y)$ location of an unobserved cell in the 1 km grid and $t^*$ is any time point (either 30-minute as is observed data or another time scale) was simulated 99 times. This simulation is run for each 1 km grid cell, yielding periodic simulations of the Fourier frequencies at each unobserved grid cell.

![Figure 6](image)

**Figure 6.** Two conditional simulations of $h$ (W/m$^2$) into the 1 km gridded surface over the SGP domain (April 29, 2005, 23:30 CST). UTM zone 14N coordinates (meters) were used to meaningfully calculate distances between sites and to create the 1 km grid.

These simulated frequencies were reconvereted back to flux values by taking the inverse DFFT and adding back in the diurnal and seasonal cycles. By sampling from the posterior distribution of the maximum likelihood estimates, uncertainty in the parameters may also be taken into account.

Figure 6 is a snapshot the spatial distribution at one time point (April 29, 2005, 23:30 CST) of 2 of the 99 simulations of $h$ over in the 1 km grid.

2.4 Computational issues and software development
The computational aspect of this project, which will be under way by the end of 2008, will deal with the data and computational complexity of this work. While we present an example analysis of sensible heat flux, many other measured parameters are of interest to climate modelers for which we will create gridded surfaces. Furthermore, there is at minimum a 10-year time series for many of these parameters, some at 1-minute intervals. The computational complexity to conduct conditional simulations of the spatio-temporal process to generate data ensembles in the 1 km grid will require both significant processing power as well as a large amount of storage. For example, at the 1 km resolution, there are ~100,000 interpolated locations and 99 simulations of the full time series (17,520 30-minute intervals per year) at each location.

To meet these challenges, we will use component-based software engineering; an approach for managing the complexity of large-scale software systems and increasing productivity. We call this software “a data domain to model domain conversion package” (DMCP).
2.5 DMCP software
The software built for this project will be capable of computing statistical descriptions of the ARM data sets on the basis of the spatio-temporal MLE described above. Furthermore, it will have the flexibility to integrate future data sets, user-defined data dependencies (such as topography and land use) and user-defined spatial and temporal scales for the gridded interpolations. Component-based software engineering (CBSE) is an emerging approach to help manage the complexity of large-scale software systems and increase the productivity of software developers and users. In CBSE, units of software are encapsulated as “components” that interact with other components only through well-defined interfaces, allowing the internal implementation of the component to remain opaque and hiding much of the complexity of the software. These more manageable units of software can be composed together to form applications. When many components are available that conform to standardized interfaces for solvers and other such capabilities, the construction of complex applications is greatly simplified, and adaptation to different problems or computational environments can be as easy as swapping one component implementation with another more suitable to the new situation. This provides a “plug-and-play” environment for the construction of applications. The US Department of Energy has supported a component approach appropriate to tackle the problems we wish to solve — the Common Component Architecture (CCA; http://www.cca-forum.org).

The CCA is a grass-roots effort to bring the benefits of component-based software engineering to high-performance scientific computing. The CCA is specifically designed to preserve the performance of components on the same processor, to support both tightly coupled parallel and distributed computing, and to simplify the incorporation of existing code into the CCA environment. The CCA uses the Babel language http://www.llnl.gov/CASC/components/ ([9] interoperability tool to allow components written in various languages (currently Fortran, C, C++, Java, and Python) to interoperate, providing the necessary support for important scientific languages and data types.

3. Conclusions and future work
The methodology for interpolating sensible heat flux data from 14 ARM EBBR sites and creating data ensembles through multiple simulations is generalizable to all parameters gathered through the ARM program. Care must be taken to pre-process the measurement data before implementing spatio-temporal interpolation and simulation, but with the development of software using CCA, data handling and computational challenges will be addressed. Initial development of the DMCP software is under way, and we expect a first version of the applications to be available by the end of 2008. Next steps also include testing the spatio-temporal process at different spatial and temporal resolutions (i.e., over 5 km or larger grid rather than 1 km grid and at longer time averages such as hourly, every 3 hours, or daily). We will also examine the multivariate processes in space and time by exploring the dependencies of land use and topography in our interpolation. This is particularly relevant for surface flux parameters, are highly dependent upon land and soil characteristics.

Acknowledgements
Work supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357 and the Argonne National Laboratory Environmental Science Division.

References
[1] Hurrell J W, Hack J J, Boville B A, Williamson D L and Kiehl J T 1998 The dynamical simulation of the NCAR Community Climate Model Version 3 J. Clim. 11:1207-1236
[2] Macduff M C and Eagan R C 2004 *ACRF data collection and processing infrastructure* (ARM Technical Report ARM TR-046)

[3] Cressie N and Huang H 1999 Classes of nonseparable spatio-temporal stationary covariance functions *J. Amer. Stat. Assoc.* **94**:1330-1340

[4] Gneiting T 2002 Nonseparable, stationary covariance functions for space-time data *J. Amer. Stat. Assoc.* **97**:590-600

[5] Stein M L 2005 Statistical methods for regular monitoring data *J. Roy. Stat. Soc. B Met.* **67**: 667-687

[6] Schneider T 2006 Analysis of incomplete data: readings from the statistics literature *Bull. Amer. Meteor. Soc.* **87**:1410-1411 [7] Hume T and Jakob C 2005 Ensemble single column modeling (ESCM) in the tropical western Pacific: Forcing data and uncertainty analysis *J. Geophys. Res.* **110**:D13109, doi:10.1029/2004JD005704

[8] Stein M L 2008 Spatial interpolation of high-frequency monitoring data *Ann. Appl. Stat.* (in press)

[9] Dahlgren T, Epperly T and Kumfert G 2003 *Babel Users' Guide, Version 0.8.4* (CASC Technical Publication, Lawrence Livermore National Laboratory)