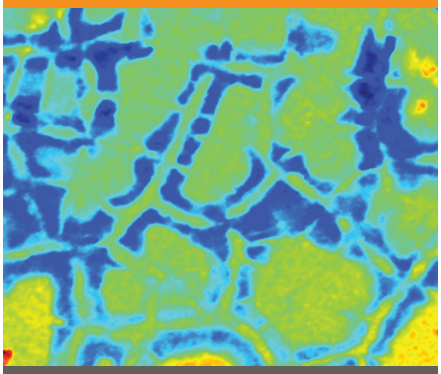
Special Section: Soil as Complex Systems



We propose a hybrid reduced-order model that efficiently predicts fine-resolution responses to forcings by first training the model with solutions from a computationally intensive model. The method was applied to Next-Generation Ecosystem Experiments–Arctic study sites to predict fine-resolution soil moisture fields based on precipitation and evapotranspiration rates. An average accuracy of 99% is achieved.

A Hybrid Reduced-Order Model of Fine-Resolution Hydrologic Simulations at a Polygonal Tundra Site

Yaning Liu, Gautam Bisht, Zachary M. Subin, William J. Riley, and George Shu Heng Pau*

High-resolution predictions of land surface hydrological dynamics are desirable for improved investigations of regional- and watershed-scale processes. Direct deterministic simulations of fine-resolution land surface variables present many challenges, including high computational cost. We therefore propose the use of reduced-order modeling techniques to facilitate emulation of fine-resolution simulations. We use an emulator, Gaussian process regression, to approximate fine-resolution four-dimensional soil moisture fields predicted using a three-dimensional surface-subsurface hydrological simulator (PFLOTRAN). A dimension-reduction technique known as "proper orthogonal decomposition" is further used to improve the efficiency of the resulting reduced-order model (ROM). The ROM reduces simulation computational demand to negligible levels compared to the underlying fine-resolution model. In addition, the ROM that we constructed is equipped with an uncertainty estimate, allowing modelers to construct a ROM consistent with uncertainty in the measured data. The ROM is also capable of constructing statistically equivalent analogs that can be used in uncertainty and sensitivity analyses. We apply the technique to four polygonal tundra sites near Barrow, Alaska that are part of the Department of Energy's Next-Generation Ecosystem Experiments (NGEE)–Arctic project. The ROM is trained for each site using simulated soil moisture from 1998–2000 and validated using the simulated data for 2002 and 2006. The average relative RMSEs of the ROMs are under 1%.

Abbreviations: ET, evapotranspiration; CESM, Community Earth System Model; DEM, Digital Elevation Model; GPR, Gaussian process regression; NGEE, Next-Generation Ecosystem Experiments; POD, proper orthogonal decomposition; PODFGPR, POD-facilitated GPR; POD-MM, POD mapping method; RRMSE, relative root mean square error; ROM, reduced-order model; SVD, singular value decomposition.

Soil moisture plays a key role in a wide range of biological and biogeochemical processes in the vadose zone. Despite its low proportion of Earth's liquid freshwater, soil moisture exerts considerable influence on rainfall–runoff relationships, land–atmosphere coupling, soil microbial diversity and activity, plant growth and physiological responses, and the need for agricultural irrigation, especially in areas with insufficient water resources (Western et al., 2002; Robinson et al., 2008).

Spatially and temporally resolved soil moisture simulations at the scale of small watersheds (100-km^2 scale), which constitute a large fraction of many landscapes, are of particular interest. In this paper we focus on the Alaskan Arctic, where a large fraction of the landscape is characterized by polygonal ground features, which are formed due to thermal expansion and contraction of ice wedges within the soil (Hinkel et al., 2005). Microtopography (meter-scale) of polygonal ground influences soil hydrologic and thermal conditions (Engstrom et al., 2005). In addition to controlling CO_2 and CH_4 emissions, soil moisture impacts (i) partitioning of incoming radiation into latent, sensible, and ground heat fluxes (Hinzman and Kane, 1992; McFadden et al., 1998); (ii) photosynthesis

Earth Sciences Division, Lawrence Berkeley National Lab., 1 Cyclotron Rd., Berkeley, CA 94720. *Corresponding author (gpau@lbl.gov).

Vadose Zone J.
doi:10.2136/vzj2014.
Received 5 May 2015.
Accepted 2 Aug. 2015.

rates (Oberbauer et al., 1991; Oechel et al., 1993; McGuire et al., 2000; Zona et al., 2011); and (iii) vegetation distributions (Gamon et al., 2012). However, performing fine-resolution simulations that explicitly resolve microtopography imposes many difficulties in practice, including computational intractability. The development of modern massively parallel supercomputing architectures and algorithms mitigates this challenge to some degree, but it is insufficient to meet the growing needs and interests of uncertainty analyses of soil moisture (Lawless et al., 2008; Harrison et al., 2012; Samaniego et al., 2013). A single or a few deterministic simulations are not sufficient, and large ensembles of experiments are necessary to provide inferential statistical properties. This "many-query" nature prevails also in the application of data assimilation, sensitivity analysis, and inverse modeling.

Typically, unsaturated flow is modeled with the highly nonlinear Richards equation (Richards, 1931; El-Kadi, 2005; Botros et al., 2012). The resulting numerical models have to be solved using Picard or Newton iterative schemes that are prone to slow or poor convergence, and in the case of Newton iterative schemes, expensive computation of derivative information (Paniconi and Putti, 1994). An approach to reduce the computational complexity of soil moisture simulations is to simplify the underlying physics and hence the governing equations, or to relax the conditions under which the model is applicable. Rodriguez-Iturbe et al. (1999) and Laio et al. (2001) developed a point (zero-dimensional) bucket model that represents soil moisture dynamics under seasonally fixed conditions by assuming that the infiltration-excess overland flow is absent. Kim et al. (1996) proposed an analytical model conditioned on the instantaneous redistribution of moisture and the linearity between evapotranspiration and soil saturation. Such models, though extremely computationally efficient, are limited in applications due to the oversimplified assumptions. Another approach that avoids direct computationally intensive simulations is through empirical or statistical modeling, usually based on remotely sensed and field-measured data (Cosby et al., 1984; Dawson et al., 1997; Hu et al., 1997; Brocca et al., 2010). In this approach, the statistical properties of soil moisture are estimated from the measured data via a regression analysis or inversion. A limitation of this approach is that the measurements can be sparse, and thereby lead to inaccurate inference.

Reduced-order models generically refer to computationally efficient mathematical or computer models that provide comparable accuracy to the computationally expensive models on which they are built. Other equivalent terms include surrogate models, emulators, response surface models, and metamodels. Reduced-order models can reduce computational cost and memory storage. There is a rich and extensive literature on ROMs (Bai and Su, 2005; Schilders et al., 2008) for a variety of applications, including hydrology (Razavi et al., 2012b). Reduced-order models facilitate analyses that necessitate a large number of model simulations for which the use of high-resolution models would be computationally impractical or impossible. Much work in model-order reduction for fine-resolution soil moisture concentrates on relating soil-moisture profiles at fine scales to those at larger spatial scales statistically (e.g., Rodriguez-Iturbe et al., 1995; Mascaro et al., 2010; Choi and Jacobs, 2011; Riley and Shen, 2014). However, the derived relationships can be complicated and depend on many factors. As such, this approach cannot be easily applied to achieve robust downscaling. Another approach is to directly include the statistical description of the spatial heterogeneity into the governing equations (Albertson and Montaldo, 2003; Montaldo and Albertson, 2003; Kumar, 2004; Teuling and Troch, 2005). This approach, however, cannot account for the temporal memory of the heterogeneity in the soil moisture, which is important for the modeling of biogeochemistry processes.

An increasingly popular approach is to use sampling-based ROMs, or emulators, as efficient surrogates for fine-scale models in these analyses (Challenor, 2012; Ratto et al., 2012). Reduced-order models are inexpensive surrogates for expensive computer simulations. They are simplifications of the model codes that aim to preserve the input–output relations in these models. Reduced-order models used in Earth system modeling are typically Bayesian statistical approaches (Sacks et al., 1989; Kennedy and O'Hagan, 2001). In particular, for Earth system models, Gaussian process regression (GPR) (Rasmussen and Williams, 2006) has typically been used in recent years for model calibration (Drignei et al., 2008; Holden et al., 2010; Bhat et al., 2010; Edwards et al., 2011; Olson et al., 2012). The calibrated emulators are then used for uncertainty quantification and sensitivity analysis (Rougier et al., 2009; Olson et al., 2012). Gaussian process regression is popular because it provides a measure of uncertainty in its estimate, allowing modelers to quantify the emulators' fidelity. However, only scalar output is being considered in these studies.

For fine-scale field solutions with a large (e.g., $\geq 10^5$) number of degrees of freedom (number of grid blocks in the discretized computational grid), ROMs that can efficiently and accurately emulate the solution of each block would be valuable. Gaussian process regression can be used directly with vectorial output by considering a covariance function that is a function of both the parameters and all components of the output (Conti and O'Hagan, 2010; Álvarez et al., 2012), allowing entire field solutions to be reconstructed. However, if we attempt to model the solutions on a fine-resolution computational grid using these techniques, the resulting GPR model will be computationally expensive. The efficiency of a GPR model can be improved using a semiparametric latent factor model (Micchelli and Pontil, 2005; Teh et al., 2005). This approach assumes that variation in the vectorial output within the parameter space can be described with significantly fewer degrees of freedom than the intrinsic number of degrees of freedom resulting from a spatial discretization. Within the theory of linear approximation space, this reduction is achieved through linear combinations of appropriate basis functions, such as eigenvectors obtained through principal component analysis

(Lawrence, 2004; Higdon et al., 2008) or wavelets (Bayarri et al., 2007; Drignei et al., 2008; Marrel et al., 2011). Gaussian process regression models are then constructed for the mixing coefficients used in the summation.

In addition to GPR, other machine-learning techniques, such as artificial neural networks (Cybenko, 1989), have been used for climate modeling (Knutti et al., 2006; Sanderson et al., 2008). For a sample of possible approaches, we refer readers to review papers by Barthelemy and Haftka (1993), Simpson et al. (2001), Lucia et al. (2004), Saridakis and Dentsoras (2008), Forrester and Keane (2009), and Razavi et al. (2012a). Of particular note are the projection-based approaches, such as the proper orthogonal decomposition (POD) method (Willcox and Peraire, 2002; Rowley et al., 2004; Cao et al., 2006; Cardoso et al., 2009; Lieberman et al., 2010), reduced basis method (Prud'homme et al., 2002; Quarteroni et al., 2011) and trajectory piecewise linearization procedure (Rewienski and White, 2003; Cardoso and Durlofsky, 2010). However, for highly nonlinear partial differential equations, projection-based approaches are difficult to apply and intrusive, requiring extensive modification to existing codes. A less intrusive approach is the POD mapping method (POD-MM) (Robinson et al., 2012; Pau et al., 2014) that uses coarse-resolution solutions to reconstruct the fine-resolution solutions. However, the computational cost of an appropriate coarse-resolution model may still be too large for some uncertainty and sensitivity analyses.

In the current work, we targeted a hybrid ROM (Higdon et al., 2008; Wilkinson, 2010) to directly emulate fine-resolution hydrological simulations and apply it to four polygonal tundra study sites at Barrow, Alaska, that are part of the Department of Energy's NGEE–Arctic project. The study of soil moisture at high latitudes is of great significance to the understanding of the driving forces of soil chemistry and greenhouse gas fluxes, and thus potential climate feedbacks of Arctic ecosystems to global climate (Schuur et al., 2008). The ROM is hybrid in the sense that we apply an emulator, namely GPR, facilitated by POD (Maxwell et al., 2007) as a dimension-reduction technique. Therefore, we obtain a direct emulated mapping from model inputs (e.g., model parameters, climate forcings) to model outputs (soil moisture, saturation, and pressure head), with large improvements in computational and memory-storage efficiencies. The average relative error of the ROM is shown to be below 1% for soil moisture, with the maximum being around 5%. Estimated confidence intervals are provided in the form of standard deviations of the Gaussian distribution, and these are demonstrated to be a reliable estimator for the approximation error in the ROM.

The paper will be organized as follows. In the Materials and Methods section, we discuss the NGEE–Arctic study sites, the fine-resolution simulations, and a detailed description of the methodology used to construct the ROM. We then describe the main results of the current work. In the final section, we summarize our findings and discuss issues that require further research.

Materials and Methods

Study Sites and Data Descriptions

The NGEE Barrow, Alaska study site (71.3° N, 156.6° W) is located on the Barrow Environment Observatory within the Arctic Coastal Plain, bounded by the Arctic Ocean to the north and the North Slope to the south. The Barrow Environment Observatory, composed of tundra, wetlands, and lakes, hosts long-term scientific investigations and environmental monitoring. The mean annual temperature and precipitation are −12.6°C and 124 mm, respectively. The surface layer of the soils is rich in organic carbon, and permafrost underlies the seasonally thawed surface layer. The four polygonal sites (labeled A, B, C, D; Fig. 1) in our study differ in microtopography. Data from a 0.25m-resolution LIDAR Digital Elevation Model (DEM) are used to characterize the microtopography at these sites. Based on the mean elevations, the four sites are categorized as low-centered (A), high-centered (B), and transitional (C and D) polygons.

PFLTORAN Simulations

For each study site, surface–subsurface isothermal flow simulations were performed with PFLOTRAN, a massively parallel reactive flow and transport model for describing subsurface processes (Hammond et al., 2014). The subsurface reactive flows and transport processes in PFLOTRAN are solved using implicit time integration and finite-volume spatial discretization. PFLOTRAN uses the Portable Extensible Toolkit for Scientific Computation (PETSc) libraries (Balay et al., 2013) for parallelization and domain decomposition. A two-dimensional diffusion-wave overland flow model was sequentially coupled with PFLOTRAN's subsurface flow model. Boundary conditions for the surface domain included precipitation and snowmelt, while evapotranspiration (ET) was applied as a sink term for the subsurface domain (i.e., evaporation from the

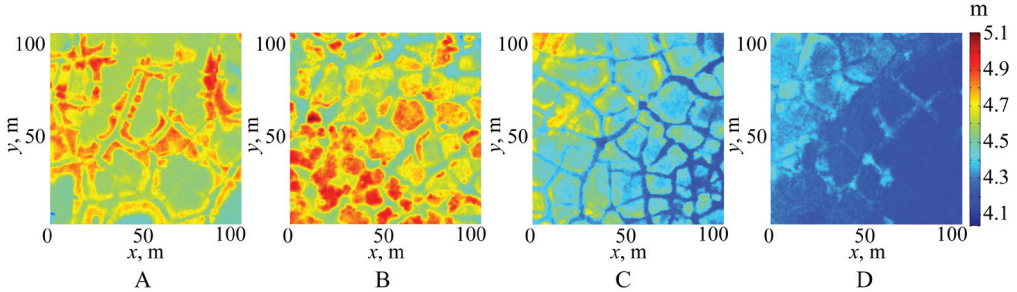


Fig. 1. The microtopography of the four sites studied under the Next-Generation Ecosystem Experiments (NGEE) Arctic project. These sites are categorized as having (A) low-centered, (B) high-centered, and (C,D) transitional polygons.

soil surface and transpiration from the root zone). The boundary conditions for PFLOTRAN were obtained by running an offline Community Land Model (CLM4.5) simulation using meteorological data (1998–2006) from the Ameriflux station in Barrow, AK. A 3000-yr CLM4.5 simulation was performed to allow for subsurface biogeochemistry in the model to reach equilibrium, and then hourly output was saved for PFLOTRAN simulations. The ET sink was distributed vertically within the PFLOTRAN subsurface domain using the same exponential rooting profile applied in CLM4.5 for Arctic shrubs (Oleson et al., 2013). In this study, we used forward simulations performed for summer months of 1998 to 2000 with PFLOTRAN surface–subsurface flows to develop ROMs for soil moisture at the four NGEE—Arctic study sites and validated the ROMs using results obtained for summer months of 2002 and 2006. The PFLOTRAN subsurface mesh is comprised of prismatic control volumes; the top triangular faces of the first layer of control volumes constitute the surface mesh. The surface mesh was terrain-following and derived from the LIDAR DEM data. The horizontal extent of the model domain is 104 × 104 m and the vertical extent is 50 cm (the approximate depth of the active layer). For a computational grid with a lateral resolution of 0.25 m and a vertical resolution of 5 cm, the total number of cells is then 3461,120. Vertical soil heterogeneity was accounted for in the PFLOTRAN simulations using soil parameter data at three depths (0–5, 5–10, and 20–25 cm) (Hinzman et al., 1991). The first two near-surface soil layers were assigned van Genuchten parameters corresponding to the 0- to 5- and 5- to 10-cm data, respectively, and the remaining eight layers were assigned soil parameters corresponding to data from 20 to 25 cm. We note that for constructing the ROM, we used solutions that are averaged onto a rectilinear grid of size 416 × 416 × 10.

ROM Development

The ROM for soil moisture was trained using the simulated soil moisture data (daily over June, July, August, and September) from 1998 to 2000 and validated using simulated soil moisture data for 2002 and 2006. Separate ROMs were constructed for each of the four study sites (A, B, C, D). The model inputs, that is, precipitation and ET, are directly mapped to the fine-resolution soil moisture at the study sites by the Gaussian process regression emulator, which is therefore distinguished from the multifidelity approach reported in Pau et al. (2014), where the input information (precipitation and ET) is not utilized. The coupling of POD alleviates the need for each grid cell in the simulation domain to develop its own emulator, whose construction can be computationally expensive.

Gaussian Process Regression (GPR)

Gaussian process regression (Rasmussen and Williams, 2006; Rasmussen and Nickisch, 2010) is a sampling-based Bayesian ROM technique that models quantities of interest as Gaussian processes conditioned on given data (observations) and prior knowledge about the input parameters. Assume that a numerical model with scalar output $f(\boldsymbol{x})$ and d-dimensional input parameters $\boldsymbol{x} = [x_1, x_2, ..., x_d]^{tr}$, where the superscript tr denotes the transpose operation, satisfies a Gaussian process (GP):

$$f(\mathbf{x}) = \mathrm{GP}[m(\mathbf{x}), k(\mathbf{x},\mathbf{x}')]$$

where $f(\mathbf{x})$ is considered as a random variable, of which any finite number of realizations are normally distributed with mean function

$$m(\mathbf{x}) = E[f(\mathbf{x})]$$

and covariance function

$$k(\mathbf{x},\mathbf{x}') = E\{[f(\mathbf{x}) - m(\mathbf{x})][f(\mathbf{x}') - m(\mathbf{x}')]\}$$

Given N pairs, $[\mathbf{x}_i, y_i = f(\mathbf{x}_i)]$, $i = 1, ..., N$, of input parameters (training set) and their corresponding model outputs (or observations), the joint distribution of the collection of known outputs $\boldsymbol{y}$ and that of the outputs $\mathbf{y}^*$ to be predicted for a set of new parameters (validation set), $\mathbf{x}_i^*$, $i = 1, ..., N^*$, follows

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} = \begin{bmatrix} m(\mathbf{X}) & k(\mathbf{X},\mathbf{X}) & k(\mathbf{X},\mathbf{X}^*) \\ m(\mathbf{X}^*)' & k(\mathbf{X}^*,\mathbf{X}) & k(\mathbf{X}^*,\mathbf{X}^*) \end{bmatrix}$$

where $\mathbf{X}$ and $\mathbf{X}^*$ are the ensembles of the training parameters $\mathbf{x}_i$ and the validation parameters $\mathbf{x}_i^*$. The posterior Gaussian distribution then follows in the simple form of

$$\mathbf{y}^*|\mathbf{X}^*,\mathbf{X},\mathbf{y} \sim \mathcal{N}\left\{k(\mathbf{X}^*,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}[f - m(\mathbf{X})] + m(\mathbf{X}^*),\right.$$
$$\left. k(\mathbf{X}^*,\mathbf{X}^*) - k(\mathbf{X}^*,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},\mathbf{X}^*)\right\} \quad [1]$$

where $\mathcal{N}(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

The choices of the mean and covariance functions are problem dependent. Without a priori knowledge of the data, we adopt in our study the constant mean function $m(\mathbf{x}) = C$ and the frequently utilized squared exponential covariance function with automatic relevance determination (covSEard) in the form of

$$k(\mathbf{X},\mathbf{X}') = \sigma_f^2 \exp\left[-\frac{1}{2}(\mathbf{X}-\mathbf{X}')^{tr}\sum{}^{-1}(\mathbf{X}-\mathbf{X}')\right] + \sigma_n^2\delta_{\mathbf{X},\mathbf{X}'}$$

Here, σ_f^2 is the variance of the model $f(\mathbf{x})$, σ_n^2 is the variance of the noise associated with the model output, $\Sigma^{-1} = \mathrm{diag}(l_1^2, l_2^2, ..., l_d^2)$ is a diagonal matrix with automatic relevance determination (ARD) parameters $l_1^2, ..., l_d^2$, and $\delta_{\mathbf{X},\mathbf{X}'}$ is the Kronecker delta function. The set of $d + 3$ hyperparameters in the mean and covariance functions, $(C, \sigma_f^2, l_i^2, i = 1, ..., d$ and $\sigma_n^2)$, are obtained by maximizing the marginal likelihood

$$\log p(\mathbf{y}\,|\,\mathbf{X}) = -\frac{1}{2}\mathbf{y}^{\mathrm{tr}}\left[k(\mathbf{X},\mathbf{X})+\sigma_n^2\mathbf{I}\right]^{-1}\mathbf{y} - \frac{1}{2}\log\left|k(\mathbf{X},\mathbf{X})+\sigma_n^2\mathbf{I}\right| - \frac{n}{2}\log 2\pi$$

where $\mathbf{I}$ is the identity matrix.

Therefore, the posterior distribution (Eq. [1]) provides a statistical estimator (the mean) for the model output values corresponding to the validation parameter set based on the given observations, while the covariance serves as an error estimator.

For multivariate output, we can theoretically construct an independent GPR approximation for each output. However, such a direct application of the scalar GPR for predicting the entire fine-resolution solution field would be computationally expensive. Even the use of multivariate Gaussian distribution (Conti and O'Hagan, 2010; Álvarez et al., 2012) is computationally challenging, given that the size of the resulting covariance matrix is proportional to the number of degrees of freedom of the multivariate output. We next describe the dimension-reduction techniques we use to address this issue.

Proper Orthogonal Decomposition

Also known as principle component analysis, the Karhunen-Loève transform, or singular value decomposition (SVD), POD (Berkooz et al., 1993; Everson and Sirovich, 1995; Kerschen et al., 2005; Maxwell et al., 2007) is a powerful statistical tool to transform a large set of correlated variables into a small number of variables that are uncorrelated. As a result, a compressed representation of the original data is obtained. The representation is optimal in the sense that the mean-squared reconstruction error is minimized.

Suppose $\mathbf{g}(\mathbf{x}) \in R^D$, where D denotes the number of degree of freedom, represents a model output field or observations. Given T samples of d-dimensional parameter $\mathbf{x}$, $\mathbf{X}_T = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, $T << D$, we can determine the corresponding solution fields $\{\mathbf{g}(\mathbf{x}_1), \mathbf{g}(\mathbf{x}_2), ..., \mathbf{g}(\mathbf{x}_T)\}$ based on the fine-resolution model. Since $\mathbf{x}$ is time-varying in our current example, $\mathbf{g}(\mathbf{x}_i)$, $i = 1, ..., T$ represent the time-series solutions of a simulation. Each field $\mathbf{g}(\mathbf{x}_i)$ is called a "snapshot" and the snapshot matrix is defined by the $D \times T$ matrix

$$S = [\mathbf{g}(\mathbf{x}^1)\ \mathbf{g}(\mathbf{x}^2) \ldots \mathbf{g}(\mathbf{x}_T)]$$

The POD approximation $\mathbf{g}^{\mathrm{POD}}$ for a given parameter takes the form of

$$\mathbf{g}^{\mathrm{POD}}(\mathbf{x}) = \overline{\mathbf{g}} + \sum_{i=1}^{M}\alpha_i\zeta_i \qquad [2]$$

where $M \leq T << D$ is the truncated dimensionality of S, $\overline{\mathbf{g}}$ is the mean snapshot

$$\overline{\mathbf{g}} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{g}(\mathbf{x}_i)$$

α_i, $i = 1, ..., M$ are the POD coefficients, and $\zeta_i = 1, ..., M$ are the POD bases, which are orthogonal to each other. The POD bases and coefficients can be determined via SVD of the snapshot matrix, or more efficiently in the case of $T << D$ through the eigenvalue decomposition of the snapshot covariance matrix defined as

$$\sum\nolimits_S = \frac{1}{T}(S-\overline{\mathbf{g}})^{\mathrm{tr}}(S-\overline{\mathbf{g}})$$

Consequently, the full set of POD bases $B^{\mathrm{F}} = [\zeta_1, \zeta_2, ..., \zeta_T]$ is in fact the left eigenvector matrix of the snapshot matrix $\mathbf{S}$ and can be obtained by normalizing $S^{\mathrm{tr}}V$, where V is the matrix of eigenvectors of the covariance Σ_S. If the model output fields are highly correlated, the magnitudes of the eigenvalues λ_i, $i = 1, ..., T$ of Σ_S decrease very rapidly, and only the M largest eigenvalues are significant. Accordingly, only a subset of the full B^{F} that correspond to the M dominant eigenvalues are included in partial set $B = [\zeta_1, \zeta_2, ..., \zeta_M]$, leading to a reduction in dimensionality. In practice, M is typically determined by

$$M = \min\left\{\bar{M}\left|1-\sum_{i=1}^{\bar{M}}\lambda_i \Big/ \sum_{i=1}^{T}\lambda_i \leq \varepsilon\right.\right\} \qquad [3]$$

where the prescribed threshold value ε represents the proportion of variance in the snapshot matrix that is not captured.

For a given parameter $\mathbf{x}$ for which $\mathbf{g}(\mathbf{x})$ is known, the best set of coefficients that minimizes $\|\mathbf{g} - \mathbf{g}^{\mathrm{POD}}\|_2$ is given by

$$\alpha_i(\mathbf{x}) = \mathbf{g}(\mathbf{x})^{\mathrm{tr}}\zeta_i \qquad i = 1, ..., M \qquad [4]$$

However, Eq. [4] cannot be used to determine the $\alpha_i(\mathbf{x})$ for which the $g(\mathbf{x})$ is unknown. Hence, we construct GPR approximations for α_i independently. We combine GPR and POD to learn and predict the direct input–output relationship and avoid the need for the intrusive modification of codes required by projection-based approach or the coarse-resolution simulations required by the POD-MM method.

POD-Facilitated GPR

The scalar GPR and POD are mutually complementary. POD reduces the dimensionality of the output fields from D to M so that it is feasible to train only M separate GPRs for the dimension-reduced problem. In the online stage, the POD coefficients are predicted with GPR and the full output space is reconstructed thereafter. We call the hybrid reduced-order model POD-facilitated GPR (PODFGPR). The whole process is summarized in Fig. 2.

To construct the GPR models for α_i, we first project the snapshots in the training set onto the POD bases to give the POD coefficients $A = [\alpha_1, \alpha_2, ..., \alpha_T] = B^{tr}S$, where each vector $\boldsymbol{\alpha}_i = [\alpha_i^1, \alpha_i^2, ..., \alpha_i^M]^{tr}$, $1 \le i \le T$ corresponds to the POD coefficients for the ith POD basis. As such, for each sample $\boldsymbol{x}_j$ in $\boldsymbol{X}_T$, we have a set of corresponding coefficients $\boldsymbol{\alpha}_i(\mathbf{x}_j)$, $i = 1, ..., M$, that can be used to perform the GPR training.

The errors associated with PODFGPR can be decomposed into two sources: the GPR approximation error, described by the posterior covariance matrix, and the POD reconstruction error due to truncating the full POD basis B^F to B, which has yet to be taken into account. Wilkinson (2010) suggested a stochastic version of POD reconstruction (Eq. [2]) by including the contribution of the ignored insignificant eigenvectors $[\zeta_{M+1}, \zeta_{M+2}, ..., \zeta_T]$ modeled as proportional to the corresponding ignored eigenvalues $[\lambda_{M+1}, \lambda_{M+2}, ..., \lambda_T]$:

$$\mathbf{f}^{\mathrm{POD}}(\mathbf{x}) = \overline{\mathbf{f}} + \sum_{i=1}^{M} \alpha_i \zeta_i + \sum_{j=1}^{T-M} \phi_j \zeta_{M+j} \qquad [5]$$

where ϕ_j, $j = 1, ..., T - M$ are independent and identically distributed (i.i.d.) zero-mean Gaussian variables with variances λ_{M+j}'s. In Eq. [5], α_i, $i = 1, ..., M$ are Gaussian predictions given by GPR (Eq. [1]), and hence the linear combination of them remains Gaussian. Combining Eq. [1] and [5], the reconstructed model output fields for a given parameter set $\mathbf{x}$ follow the distribution

$$\mathbf{f}^{\mathrm{POD}}(\mathbf{x}) \sim \mathcal{N}\left(\overline{\mathbf{f}} + \sum_{i=1}^{M}\left\{k_i(\mathbf{x},\mathbf{X})k_i(\mathbf{X},\mathbf{X})^{-1}\left[\boldsymbol{\alpha}_i^t - \mathbf{m}_i(\mathbf{X}) + m_i(\mathbf{X})\right]\right\}\zeta_i,\right.$$
$$\left.\sum_{i=1}^{M}\left[k_i(\mathbf{x},\mathbf{x}) - k_i(\mathbf{x},\mathbf{X})k_i(\mathbf{X},\mathbf{X})^{-1}k_i(\mathbf{X},\mathbf{x})\right]\zeta_i^2 + \sum_{j=1}^{T-M}\lambda_{M+j}\zeta_{M+j}^2\right) \qquad [6]$$

where $\boldsymbol{\alpha}_i^t = \boldsymbol{\alpha}_i^t(\mathbf{X}) = [\alpha_i^t(\mathbf{x}_1), \alpha_i^t(\mathbf{x}_2), ..., \alpha_i^t(\mathbf{x}_N)]^{tr}$ is the vector of the ith POD coefficients associated with all the N training samples, $k_i(\cdot,\cdot)$, $\mathbf{k}_i(\cdot,\cdot)$, $m_i(\cdot)$, $\mathbf{m}_i(\cdot)$ denote the covariance(s) and mean(s) of the ith POD coefficient evaluated at the given parameter set(s). Thus, we have developed a posterior Gaussian formulation for the model output fields. The mean of the Gaussian posterior serves as the most likely prediction for the solution field based on the training data. In addition, the standard deviation in the Gaussian probabilistic prediction provides an estimator for the error of PODFGPR for a given credibility level. For example, the mean predictions minus and plus two standard deviations mark the range the true values are expected to fall in with a probability of around 95%, and therefore can be used as a confident characterization of the uncertainty associated with the predictor.

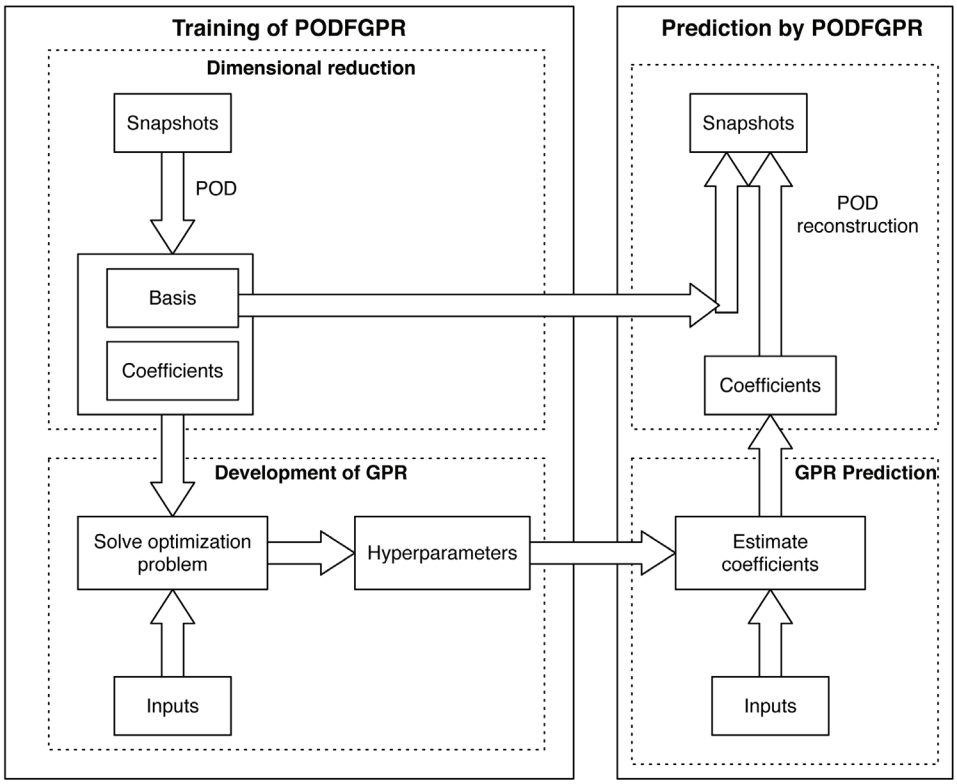


Fig. 2. Demonstration of proper orthogonal decomposition facilitated Gaussian process regression (PODFGPR).

Results

We constructed PODFGPR models for the simulated 0.25-m-resolution four-dimensional summer soil moisture data for the four NGEE—Arctic Barrow sites. With minimal loss of generality, we only present the results for Sites A and D, for which the levels of spatial heterogeneity are most different, as reflected by the numbers of POD bases needed to capture the same percentage of variability.

Analysis of the Parameter Space

The soil moisture data are divided into two parts for cross validation. The 1998–2000 data are used for training purposes. The resulting ROM is then validated using the 2002 and 2006 data. The model input parameters considered in this study are ET rate (kg s^{-1}) and precipitation rate (m^3 s^{-1}) obtained from CLM4.5 simulation that was performed using meteorological data from the Ameriflux station in Barrow, AK (Fig. 3, shown for summer months). Both input parameters are horizontally homogeneous. Thus, the effective number of model input parameters in this study is two: ET and precipitation rate.

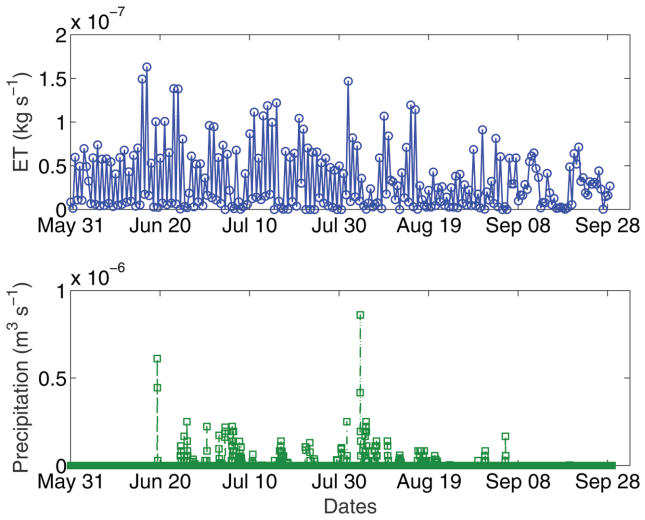


Fig. 3. Time series of top-layer (0–5 cm) evapotranspiration (ET) and precipitation rates of 2000 at the interval of 1 and 12 h, respectively.

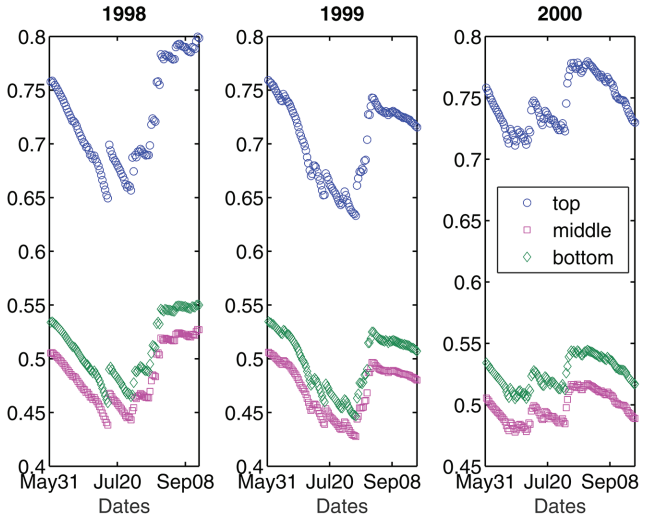


Fig. 4. Daily site A soil moisture time series of three 0.25-m grid cells at three depths (0–5, 20–25, and 45–50 cm). The fluctuation patterns are similar for different layers of the same year, as well as of different years.

Figure 3 shows the time series of the top-layer ET and precipitation rates for the summer months of the year 2000. The precipitation rate increases on June 20 and remains relatively high until September 5, after which it drops. The ET rate typically increases following precipitation. The input data of the other years also show similar behaviors.

The soil moisture simulation data are recorded once per day. For Sites A and D, data are recorded for 110 and 80 d, respectively, for the year of 1998. In other instances, data are recorded for 120 d of the year. Figure 4 plots the daily time series of soil moisture for three arbitrarily chosen spatial points located at the top (0–5 cm), middle (20–25 cm), and bottom (45–50 cm) soil layers of Site A. For each year, all the layers share a similar pattern in the soil moisture change. For 1998, the moisture content declines until July 13, then increases following a precipitation event, decreases until July 30, and increases again following precipitation on July 31. The years 1999 and 2000 are atypical in that the soil moisture goes down after August 20 due to the low precipitation.

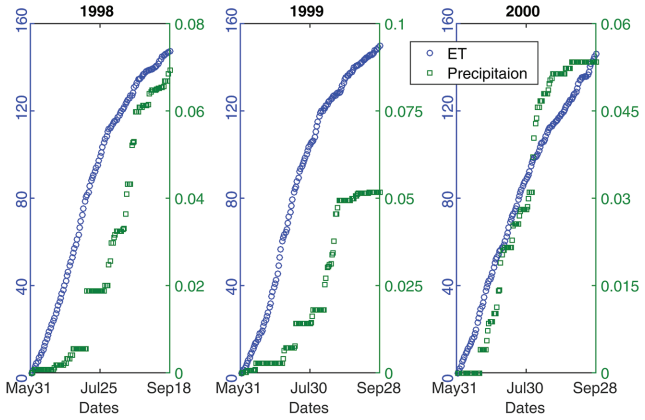


Fig. 5. Normalized accumulated evapotranspiration (ET) (left *y* axis) and accumulated precipitation (m^3) (right *y* axis) for the training years that are used as the input parameters to construct the reduced-order model (ROM). The highest and lowest accumulated ET happen in 1999 and 2000, and 1998 and 1999 for accumulated precipitation.

To match the hourly input parameters with the daily output data, the ET and precipitation rates are integrated for each day to obtain the daily values. The accumulated daily values are used as input parameters in the construction of the ROM (Fig. 5). The year 1999 has the highest accumulated ET and lowest accumulated precipitation values. The highest accumulated precipitation value happens in 1998 and the lowest accumulated ET value happens in 2000.

Convergence of the POD

For each site, POD is performed for the training snapshots (soil moisture fields for the years 1998–2000). Figure 6 shows that the magnitudes of the eigenvalues of the snapshot matrices for Sites A and D decrease fairly quickly (blue solid lines marked with circles). Also plotted is the noncaptured fraction of the variance used to determine the number of POD basis in Eq. [3]. Setting $\varepsilon = 10^{-6}$, 8 and 24 bases are selected for Sites A and site D, respectively. Throughout the following text, we will use these numbers in the POD construction and reconstruction.

Figure 7 displays the performance of POD for the predicted moisture fields of Sites A and D on 30 July 2006 (date arbitrarily chosen). Column 1 shows the PFLOTRAN-simulated soil moisture $\mathbf{f}$ at the middle layer. Projecting the snapshots on the POD bases generates the POD coefficients $\{\alpha_i\}_{i=1}^{M}$, which are used to reconstruct the POD approximations $\mathbf{f}^{\mathrm{POD}}$ by Eq. [2] (Column 2). Column 3 shows the pointwise relative errors $|(\mathbf{f} - \mathbf{f}^{\mathrm{POD}})/\mathbf{f}|$ at each grid cell. The errors are on the order (notated here as "*O*")

$O(10^{-6})$ – $O(10^{-4})$. Note that the POD reconstructed snapshot $\mathbf{f}^{POD}$ does not serve as a practical ROM, as the "true" solution field $\mathbf{f}$ is not known in the first place. Rather, the relative error shown in this pure POD analysis provides a lower bound on the approximation error of PODFGPR, achievable only if there are no approximation errors in the GPR models of the POD coefficients.

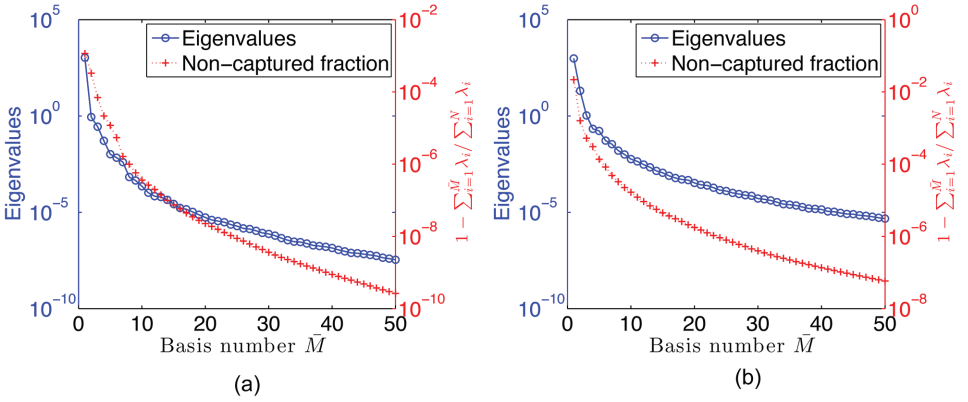


Fig. 6. Eigenvalues (left *y* axis) and the noncaptured fraction of the variability (right *y* axis) of the training data for (a) Site A and (b) Site D. Both sites show rapid rates of decrease of eigenvalues. Site D shows a slower rate compared to Site A.

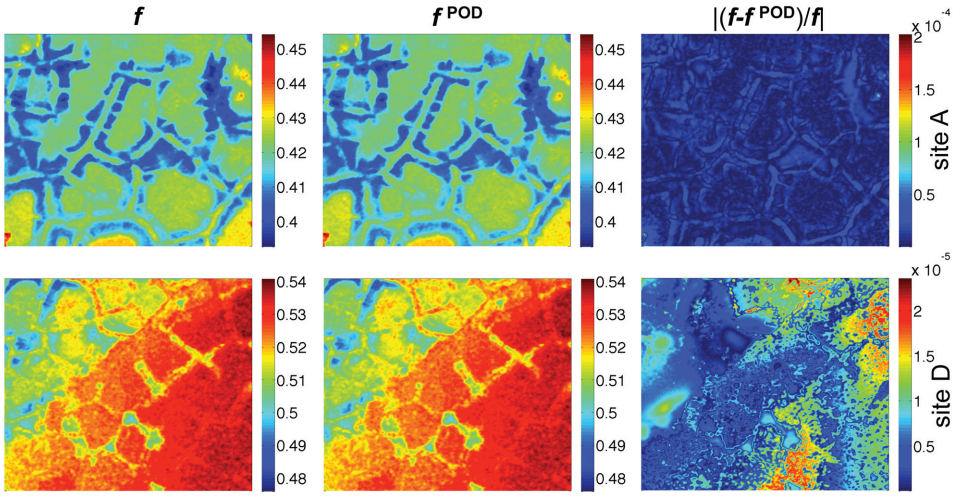


Fig. 7. Proper orthogonal decomposition (POD) for the 20- to 25-cm layer moisture fields of Sites A (row 1) and D (row 2) on 30 July 2002. Column 1: the simulated fields *f*, Column 2: the POD-reconstructed fields $\mathbf{f}^{POD}$, Column 3: the pointwise relative errors $|(\mathbf{f}-\mathbf{f}^{POD})/\mathbf{f}|$. The moisture fields are captured by the POD reconstruction with errors < 10^{-4}, which are due to the truncation of the POD basis set.

Performance of PODFGPR

Next we analyze the performance of GPR in learning and predicting the input–output relationship between the input parameters and the POD coefficients for sites A and D (Fig. 8). In the GPR construction, the training input parameters are scaled to the unit interval [0, 1], and the training POD coefficients, computed by projecting the training snapshots to the POD bases through Eq. [4], are normalized by their standard deviations so that they have unit variances. In the prediction stage, the POD coefficients corresponding to the validation snapshots, α^{GPR}, are predicted based on Eq. [1] and denormalized by the standard deviations in the training POD coefficients. Since the validation snapshots are known in this case, the true POD coefficients, α^{POD}, can be computed based on Eq. [4]. We only show the time series of α_1 since the first POD basis describes most of the variance in the training snapshot matrix (Fig. 6). Hence, the capability of capturing the first POD coefficients throughout the prediction period is crucial to the overall performance of the final ROM. The two sites follow a similar pattern in the change of α_1. The predicted and true values (α^{GPR} and α^{POD}) are in good agreement. The errors are mostly bounded by the two-standard-deviation bounds, computed from the covariance matrix in Eq. [1]. The time series of other α^{GPR} and α^{POD} were also compared and their differences are consistently small, falling between the two-standard-deviation bounds.

We now examine the overall performance of PODFGPR. Since the POD basis $\{\zeta_i\}$, $i = 1, ..., M$ and α^{GPR} for the input parameters of the years 2002 and 2006 have been calculated, the PODFGPR-based ROM predictions for the soil moisture fields can be readily computed as the mean of the Gaussian distribution given by Eq. [6]. Figure 9 compares the simulated soil moisture fields ($\mathbf{f}$) on 12 July 2006 at Site A and on 6 July 2002 at site D with the PODFGPR predicted fields ($\mathbf{f}^{PODFGPR}$) at three different layers. The pointwise relative errors ($|\mathbf{f}-\mathbf{f}^{PODFGPR}/\mathbf{f}|$), which are of order $O(10^{-3})$, are plotted in the third column. In Site D, the ROM error exhibits a similar pattern to the simulated moisture, indicating that there exists a positive correlation between them. Both the ROM errors and the simulated moisture magnitudes are higher in the east and southeast corner, but are low in the northwest corner. We note that since the emulation of α^{POD} is not physically constrained, $\mathbf{f}^{PODFGPR}$ may have values that are nonphysical. This result is typically obtained if a large number of POD bases are used. The POD bases associated with smaller eigenvalues tend to be more oscillatory and failure to approximate the associated POD coefficients accurately can lead to large fluctuations that lead to unphysical

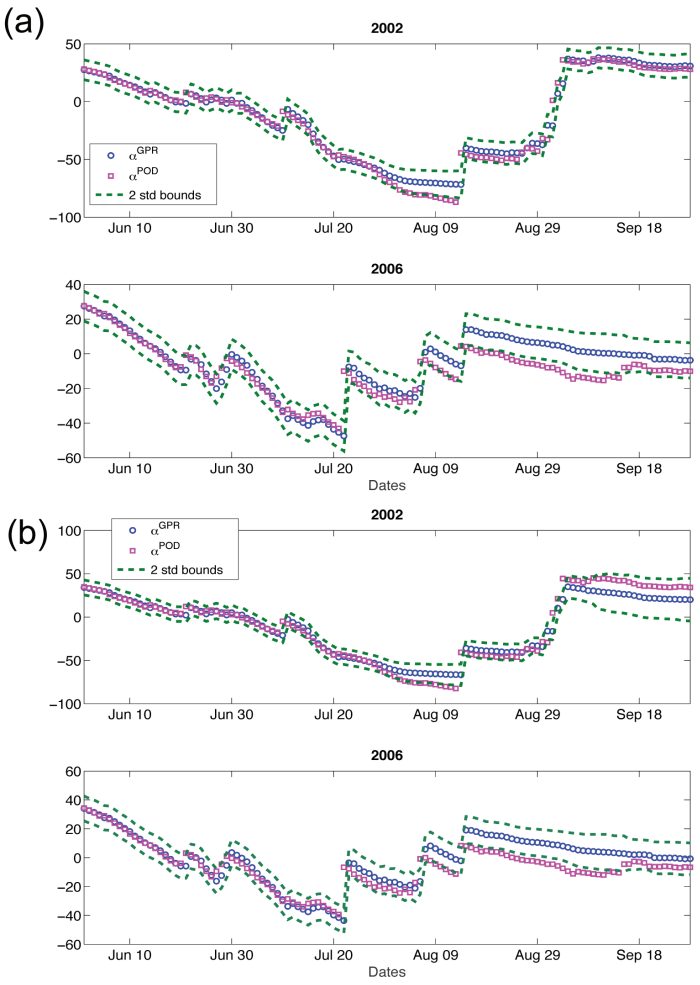


Fig. 8. Gaussian process regression (GPR) predictions (α^{GPR}) and two-standard-deviation error bounds (2 std bounds) for the time series of the first proper orthogonal decomposition (POD) coefficients of the 2002 and 2006 snapshots (α^{POD}) at (a) Site A and (b) Site D. The α^{POD} is within 2 standard deviations from α^{GPR} in most cases.

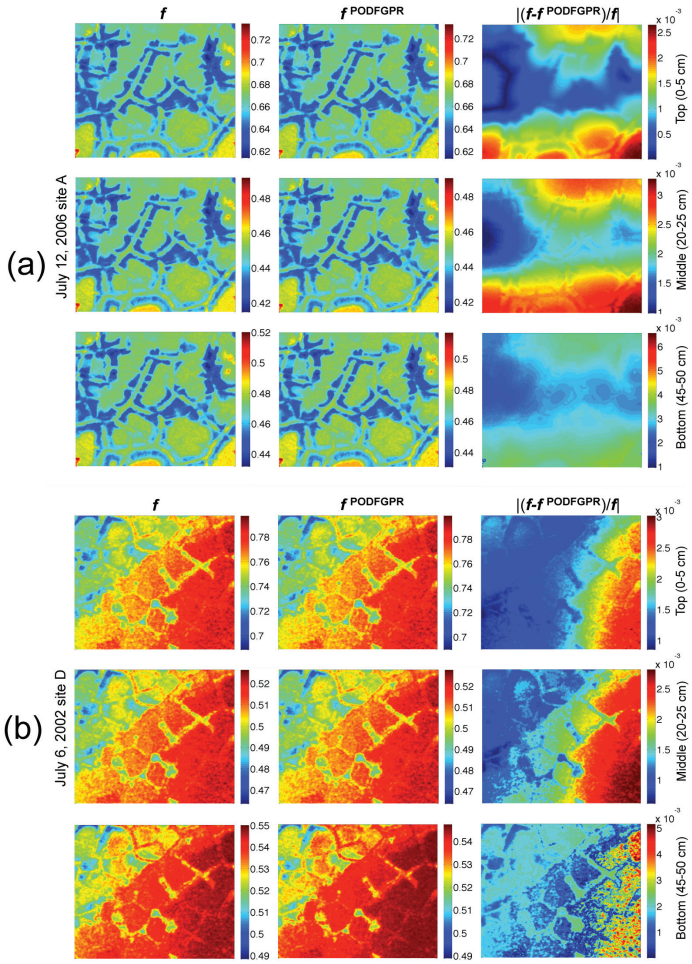


Fig. 9. Comparison of PFLOTRAN simulated soil moisture fields f (left column) with PODFGPR predictions $f^{PODFGPR}$ (middle column) at the top, middle, and bottom layers on 12 July 2006 at Site A (a) and on 6 July 2002 at Site D (b). The right column is the pointwise relative errors $|(\mathbf{f}-\mathbf{f}^{PODFGPR})/\mathbf{f}|$. The simulated moisture fields are accurately predicted by the reduced-order model (ROM) with errors of $O(10^{-3})$.

results. In the current work, we ensured that the number of POD bases used does not lead to unphysical results in the validation sample set.

To further quantify the approximation error, we use the relative root mean square errors (RRMSE) to measure the overall accuracy for the full soil moisture field. The RRMSE associated with an arbitrary time (day in our example) i, RRMSE_i, is defined as

$$\mathrm{RRMSE}_i = \sqrt{\sum_{j=1}^{D}\left(\frac{f_{i,j}-f_{i,j}^{\mathrm{PODFGPR}}}{f_{i,j}}\right)^2 \Big/ D}$$

where $\{f_{i,j}\}, j = 1, ..., D$ comprise $\mathbf{f}_i$, the snapshot at time i obtained by simulating PFLOTRAN and similarly, $\{f_{i,j}^{\mathrm{PODFGPR}}\}, j = 1, ..., D$ are the elements of the PODFGPR-predicted snapshot.

Figure 10 compares the RRMSEs of POD and PODFGPR, as well as their ratios for all the validation samples of Sites A and D. On average, the error in PODFGPR is 242 times greater than that of POD, indicating that the GPR prediction errors in the PODFGPR-based ROM lead to two orders of magnitude reduction in accuracy for the current study. Figure 11a presents the histogram of the RRMSEs for the combined 480 snapshots in the two validation years at Sites A and D. The average of all the errors is 0.86%. Figure 11b shows the (empirical) cumulated frequencies of the RMMSE for three cases: all the validation data (including both Sites A and D), validation data only related to Site A, and validation data only related to Site D. For all the data, 67.29% of the errors are below 1%, and more than 90% are below 2%. The ROM also performs slightly better for Site A, as its cumulated frequency curve is above that of Site D for the most part. We conclude that the PODFGPR ROM demonstrates good accuracy for predicting soil moisture of

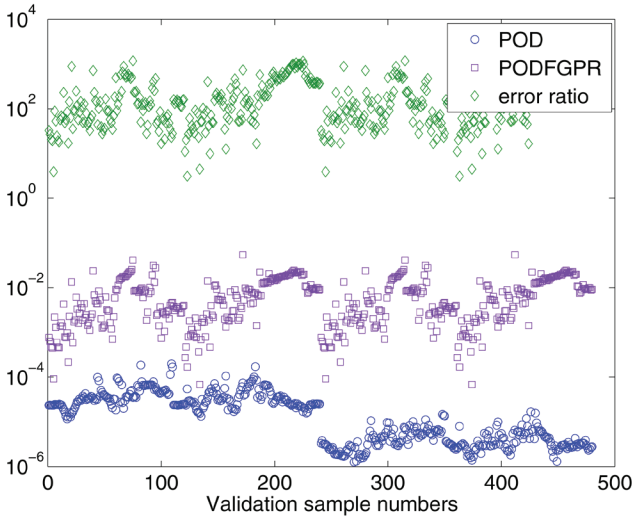


Fig. 10. Relative root mean square errors (logarithmic scale) of POD and PODFGPR, as well as their ratios (error ratio), for the combined samples of Sites A (first 240) and D (last 240). The mean of the ratios is 242, showing that proper orthogonal decomposition (POD) is two orders of magnitude more accurate than PODFGPR.

the entire validation period, taking into consideration the inherent uncertainty in the fine-resolution model.

The computational cost of constructing the PODFGPR ROM and predicting soil moisture fields is negligible once the training snapshots are available. The construction process in the considered example takes approximately 10 min on a single 2.3-GHz CPU and predicting a snapshot takes a few seconds. In contrast, one forward simulation with PFLOTRAN for a period of 120 d requires 24 h of wall-clock time with 96 processors (2304 CPU hours).

Compared with the POD-MM discussed in Pau et al. (2014), the PODFGPR ROM is one order of magnitude less accurate for this particular problem. This error budget possibly indicates that the nature of the problem determines that it is easier for the ROM to assimilate the information contained in the coarse-resolution solutions than to learn from the input parameter values (the construction of ROM in this context can be viewed as a machine-learning problem). However, the PODFGPR approach avoids the cost of additional coarse-resolution simulations that drive POD-MM. In real-world applications, it is often the case that meaningful coarse-resolution simulations also have nonnegligible computational cost. Moreover, PODFGPR can more easily provide a framework for developing a site-independent ROM. By taking site-dependent parameters, such as topography, into account, PODFGPR can provide approximation of the soil moisture at any arbitrary site without the need to create a coarse model. Creating coarse models for multiple sites can quickly become nontrivial.

Uncertainty Estimate of the ROM

The PODFGPR-based ROM provides an uncertainty estimator in the form of the Gaussian standard deviation (Eq. [6]), which is useful in the context of uncertainty analysis (e.g., in Monte Carlo approaches). The use of a ROM enables a large number of ensembles of fine-resolution simulations, but at the same time entails an additional source of uncertainty that can be accounted for in interpreting the results. Many ROMs lack the ability to quantify their own uncertainty and are used as if none existed. While the uncertainty of the ROMs can be estimated from cross-validation, it requires more fine-scale simulations and only provides an averaged parameter-independent uncertainty estimate. In contrast, the PODFGPR-based ROM provides a parameter-dependent uncertainty estimate based solely on the training data.

We examine the behaviors of the standard deviation bounds in Fig. 12. The (absolute) RMSEs, defined in a similar way to RRMSE but without normalization by the simulated snapshots, are plotted for all the validation data at Sites A and D. The one-, two-, and three-standard-deviation bounds in the RMSE sense are also charted side by side. The bounds are capable of precisely capturing the increase in errors on approximately July 30 and September 3 in 2002, as well as the one on August 19 in 2006, but with a lag of 10 d. The two standard deviations can be regarded as a reliable bound, encompassing a majority of the RMSEs. Combining Sites A and D and years 2002 and 2006, 50.83%, 78.12%, and 95.63% of the RMSEs are bounded by the one-, two-, and three-standard deviation bounds, respectively.

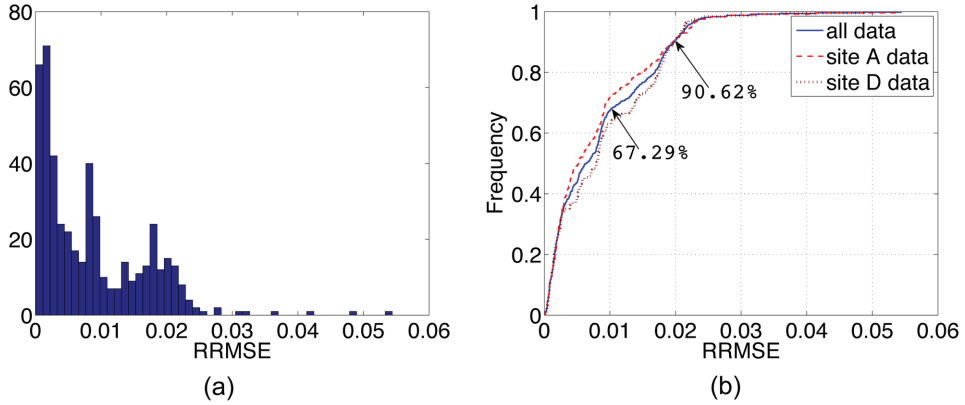


Fig. 11. Histogram (a) and empirical cumulative frequency plots (b) of the relative root-mean-square errors demonstrating the accuracy of the reduced-order model (ROM). More than 90% of the errors are under 2%.

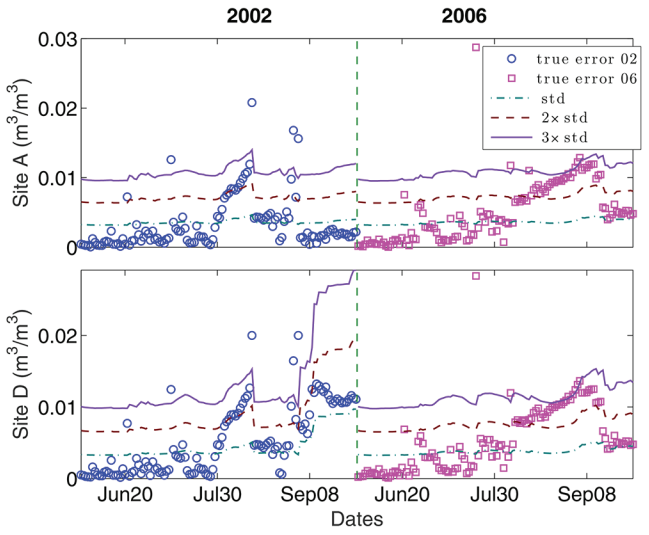


Fig. 12. Absolute root mean square errors of proper orthogonal decomposition facilitated Gaussian process regression (PODFGPR) of 2002 and 2006 (true error 02 and true error 06) at Sites A and D, as well as the one-, two-, and three-standard-deviation bounds (std, 2×std, and 3×std). The vertical green dashed lines separate the results of 2002 from those of 2006. The bounds serve as dependable indicators (confidence intervals) of the true errors of the reduced-order model (ROM).

Discussion

A limitation of the approach presented herein is that the ROM is site-specific. It would be valuable to construct a site-independent ROM by including the DEM data. One approach is to treat the moisture value of each computational grid cell as a sample, leading to an extremely large sample pool. The computational complexity is then similar to that of a multivariate GPR. A possible solution is to implement an adaptive sampling algorithm (Pau et al., 2013) that selects a subset of representative samples and significantly reduces the effect of overfitting. A potential problem with this approach is that, in the case where a small group of adaptively selected samples are insufficient to produce an accurate emulator, it will eventually be as costly as regular GPR with more samples included, since each sample selection involves an offline GPR procedure for the selected parameter group and online predictions for the rest. Another approach to constructing a site-independent ROM is to parameterize the two-dimensional topography of all sites for which the ROM is built, for examply, by using a Karhunen–Loève decomposition procedure. However, simulations must be performed on many of these sites to have sufficient data to construct a ROM that can be robustly used at all the sites.

The uncertain parameters considered in this paper are only precipitation and ET. If additional parameters were considered, it could increase the complexity of the PODFGPR by increasing the variability in the snapshots, calling for more complex GPR models that can accurately map the input parameters to the POD coefficients, or necessitating more training samples to capture the response of parameters in an enlarged parameter space. Most sampling-based reduced-order modeling techniques rely on their abilities to identify simple relationships between the parameters and the output that are not a priori known. Mathematically, reduced-order models attempt to find the low-dimensional parameter-induced manifold in the high-dimensional discrete space in which the solutions lie (Cuong et al., 2005). If a problem is too complex for such a manifold to exist, then a reduced-order model cannot be built efficiently. The approach described in this work provides a systematic approach to achieving this task, and adapting to more complex relationships, for example, by employing more complex GPR models.

A potential way to improve the straightforward GPR model in the current work is to treat the POD coefficients as a dynamically evolving vectorial Gaussian process, such as by modeling the POD coefficients as a first-order Markov chain. Under this assumption, the POD coefficients of the current day, in our example, are only dependent on the POD coefficients of the previous day and the daily-averaged or integrated precipitation and ET fluxes of the current day.

The PODFGPR-based ROM can be employed as the replacement of the original high-fidelity models to facilitate Monte Carlo–based analyses, such as variance-based global sensitivity analysis and uncertainty quantification, along with efficient sampling strategies (e.g., Liu et al., 2013, 2015). In these analyses, the estimated standard deviation in the ROM may have nonnegligible impacts on the results when the ROM is used as a surrogate for high-fidelity simulations. Marrel et al. (2011) showed that taking the standard deviation into consideration leads to Sobol' indices that are more robust than using only the predicted output. In addition, a confidence interval can be constructed for these indices. By quantifying the relative importance of variance in these analyses, we can construct a ROM with uncertainty that matches the uncertainty in the observations and thus improve the efficiency of the ROM.

The ROM developed in this work can potentially be used to approximate a multiscale model of a component of the Community Earth System Model (CESM). Instead of constructing a single ROM for the land model, we construct a ROM for each of the processes or submodels in a CESM component. This hierarchy of ROMs allows us to avoid directly modeling the complex responses resulting from interactions between processes. However, in a multiscale setting, coupling between two submodels frequently involves upscaling or downscaling of a coupling variable that is vectorial in nature. When these two submodels are ROMs, the coupling variable can be appropriately parameterized based on linear approximation theory (represented by the POD coefficients), and the input–output relations modeled by the ROMs can be constructed for this reduced representation. We would thus avoid the vectorial reconstruction of the coupling variable—a cost savings that will be significant if the dimension of the coupling variable is large. This approach has the potential of realizing a hierarchical

ROM with a computational complexity that is independent of the dimension of the underlying discretization.

Conclusions

In this paper, we presented an efficient hybrid reduced-order model that combines the Gaussian process regression emulator and the dimension-reduction tool proper orthogonal decomposition. The advantage of this approach over the POD-MM method described in Pau et al. (2014) is that the coarse-resolution simulations are circumvented by directly modeling the input-output relationship of the fine-resolution hydrological model. We apply the ROM to emulate the soil moisture at NGEE—Arctic study sites and achieve an average RRMSE of less than 1%. In addition, the ROM is equipped with a desired estimator of prediction error in the form of a Gaussian standard deviation, which can be utilized in uncertainty analysis. Scalable high-performance software that can perform GPR and POD for large datasets is under development to handle problems with a large number of degrees of freedom and snapshot samples. We will explore such applications in future work.

Acknowledgments

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the US Department of Energy under Contract #DEAC02-05CH11231 as part of the Early Career Research Program (Liu and Pau) and the Terrestrial Ecosystem Science Program, including the Next-Generation Ecosystem Experiments (NGEE—Arctic) project (Bisht and Riley). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under the aforementioned contract.

References

Albertson, J.D., and N. Montaldo. 2003. Temporal dynamics of soil moisture variability: 1. Theoretical basis. Water Resour. Res. 39(10). doi:10.1029/2002WR001616

Álvarez, M. A., L. Rosasco, and N. D. Lawrence. 2012. Kernels for vector-valued functions: A review. Found. Trends Mach. Learn. 4:195–266.

Bai, Z.J., and Y.F. Su. 2005. Dimension reduction of large-scale second-order dynamical systems via a second-order Arnoldi method. SIAM J. Sci. Comput. 26:1692–1709. doi:10.1137/040605552

Balay, S., J. Brown, K. Buschelman, W.D. Gropp, D. Kaushik, M.G. Knepley, L.C. McInnes, B.F. Smith, and H. Zhang. 2013. PETSc users manual. Argonne National Lab., Lemont, IL.

Barthelemy, J.F.M., and R.T. Haftka. 1993. Approximation concepts for optimum structural design—A review. Struct. Multidiscip. Optim. 5:129–144. doi:10.1007/BF01743349

Bayarri, M.J., J.O. Berger, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R.J. Parthasarathy, R. Paulo, J. Sacks, and D. Walsh. 2007. Computer model validation with functional output. Ann. Stat. 35:1874–1906. doi:10.1214/009053607000000163

Berkooz, G., P. Holmes, and J.L. Lumley. 1993. The proper orthogonal decomposition in the analysis of turbulent flows. Annu. Rev. Fluid Mech. 25:539–575. doi:10.1146/annurev.fl.25.010193.002543

Bhat, K.S., M. Haran, and M. Goes. 2010. Computer model calibration with multivariate spatial output: A case study. In: M-H. Chen et al., editors, Frontiers of statistical decision making and Bayesian analysis. Springer, New York. p. 168–184.

Botros, F.E., Y.S. Onsoy, T.R. Ginn, and T. Harter. 2012. Richards Equation-based modeling to estimate flow and nitrate transport in a deep alluvial vadose zone. Vadose Zone J. 11(4). doi:10.2136/vzj2011.0145

Brocca, L., F. Melone, T. Moramarco, and R. Morbidelli. 2010. Spatial-temporal variability of soil moisture and its estimation across scales. Water Resour. Res. 46. doi:10.1029/2009WR008016

Cao, Y., J. Zhu, Z. Luo, and I.M. Navon. 2006. Reduced-order modeling of the upper tropical Pacific ocean model using proper orthogonal decomposition. Comput. Math. Appl. 52:1373–1386. doi:10.1016/j.camwa.2006.11.012

Cardoso, M.A., and L.J. Durlofsky. 2010. Linearized reduced-order models for subsurface flow simulation. J. Comput. Phys. 229:681–700. doi:10.1016/j.jcp.2009.10.004

Cardoso, M.A., L.J. Durlofsky, and P. Sarma. 2009. Development and application of reduced-order modeling procedures for subsurface flow simulation. Int. J. Numer. Methods Eng. 77:1322–1350. doi:10.1002/nme.2453

Challenor, P. 2012. Using emulators to estimate uncertainty in complex models. In: Andrew M. Dienstfrey and Ronald F. Boisvert, editors, Uncertainty quantification in scientific computing. IFIP Advances in Information and Communication Technology 377. Springer, New York. p. 151–164.

Choi, M., and J.M. Jacobs. 2011. Spatial soil moisture scaling structure during Soil Moisture Experiment 2005. Hydrol. Processes 25:926–932. doi:10.1002/hyp.7877

Conti, S., and A. O'Hagan. 2010. Bayesian emulation of complex multi-output and dynamic computer models. J. Stat. Plan. Inference 140:640–651. doi:10.1016/j.jspi.2009.08.006

Cosby, B.J., G.M. Hornberger, R.B. Clapp, and T.R. Ginn. 1984. A statistical exploration of the relationships of soil-moisture characteristics to the physical-properties of soils. Water Resour. Res. 20:682–690. doi:10.1029/WR020i006p00682

Cuong, N.N., K. Veroy, and A.T. Patera. 2005. Certified real-time solution of parametrized partial differential equations. In: S. Yip, editor, Handbook of materials modeling. Springer, New York. p. 1523–1558.

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. Math. Contr. Signals Syst. 2:303–314. doi:10.1007/BF02551274

Dawson, M.S., A.K. Fung, and M.T. Manry. 1997. A robust statistical-based estimator for soil moisture retrieval from radar measurements. IEEE Trans. Geosci. Rem. Sens. 35:57–67. doi:10.1109/36.551934

Drignei, D., C.E. Forest, and D. Nychka. 2008. Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling. Ann. Appl. Stat. 2:1217–1230. doi:10.1214/08-AOAS210

Edwards, N.R., D. Cameron, and J. Rougier. 2011. Precalibrating an intermediate complexity climate model. Clim. Dyn. 37:1469–1482. doi:10.1007/s00382-010-0921-0

El-Kadi, A. 2005. Validity of the generalized Richards equation for the analysis of pumping test data for a coarse-material aquifer. Vadose Zone J. 4:196–205. doi:10.2136/vzj2005.0196

Engstrom, R., A. Hope, H. Kwon, D. Stow, and D. Zamolodchikov. 2005. Spatial distribution of near surface soil moisture and its relationship to microtopography in the Alaskan Arctic coastal plain. Nord. Hydrol. 36:219–234.

Everson, R., and L. Sirovich 1995. Karhunen-Loeve Procedure for gappy data. J. Opt. Soc. Am. 12:1657–1664. doi:10.1364/JOSAA.12.001657

Forrester, A.I.J., and A.J. Keane. 2009. Recent advances in surrogate-based optimization. Prog. Aerosp. Sci. 45:50–79. doi:10.1016/j.paerosci.2008.11.001

Gamon, J.A., G.P. Kershaw, S. Williamson, and D.S. Hik. 2012. Microtopographic patterns in an arctic baydjarakh field: Do fine-grain patterns enforce landscape stability. Environ. Res. Lett. 7(1). doi:10.1088/1748-9326/7/1/015502

Hammond, G.E., P.C. Lichtner, and R.T. Mills. 2014. Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLOTRAN. Water Resour. Res. 50:208–228. doi:10.1002/2012WR013483

Harrison, K.W., S.V. Kumar, C.D. Peters-Lidard, and J.A. Santanello. 2012. Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques. Water Resour. Res. 48. doi:10.1029/2012WR012337

Higdon, D., J. Gattiker, B. Williams, and M. Rightley. 2008. Computer model calibration using high-dimensional output. J. Am. Stat. Assoc. 103:570–583. doi:10.1198/016214507000000888

Hinkel, K.M., R.C. Frohn, F.E. Nelson, W.R. Eisner, and R.A. Beck. 2005. Morphometric and spatial analysis of thaw lakes and drained thaw lake basins in the western Arctic Coastal Plain, Alaska. Permafrost Periglacial Processes 16:327–341. doi:10.1002/ppp.532

Hinzman, L.D., and D.L. Kane. 1992. Potential Response of an Arctic Watershed during a Period of Global Warming. J. Geophys. Res., D, Atmospheres 97(D3):2811–2820. doi:10.1029/91JD01752

Hinzman, L.D., D.L. Kane, R.E. Gieck, and K.R. Everett. 1991. Hydrologic and thermal-properties of the active layer in the Alaskan arctic. Cold Reg. Sci. Technol. 19:95–110. doi:10.1016/0165-232X(91)90001-W

Holden, P.B., N.R. Edwards, K.I.C. Oliver, T.M. Lenton, and R.D. Wilkinson. 2010. A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. Clim. Dyn. 35:785–806. doi:10.1007/s00382-009-0630-8

Hu, Z.L., S. Islam, and Y.Z. Cheng. 1997. Statistical characterization of remotely sensed soil moisture images. Remote Sens. Environ. 61:310–318. doi:10.1016/S0034-4257(97)89498-9

Kennedy, M.C., and A. O'Hagan. 2001. Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B Stat. Methodol. 63:425–450. doi:10.1111/1467-9868.00294

Kerschen, G., J.C. Golinval, A.F. Vakakis, and L.A. Bergman. 2005. The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: An overview. Nonlinear Dyn. 41:147–169. doi:10.1007/s11071-005-2803-2

Kim, C.P., J.N.M. Stricker, and P.J.J.F. Torfs. 1996. An analytical framework for the water budget of the unsaturated zone. Water Resour. Res. 32:3475–3484. doi:10.1029/95WR02667

Knutti, R., G.A. Meehl, and M.R. Allen. 2006. Constraining climate sensitivity from the seasonal cycle in surface temperature. J. Clim. 19:4224–4233. doi:10.1175/JCLI3865.1

Kumar, P. 2004. Layer averaged Richard's equation with lateral flow. Adv. Water Resour. 27:521–531. doi:10.1016/j.advwatres.2004.02.007

Laio, F., A. Porporato, L. Ridolfi, and I. Rodriguez-Iturbe. 2001. Plants in water-controlled ecosystems: Active role in hydrologic processes and response to water stress. II. Probabilistic soil moisture dynamics. Adv. Water Resour. 24:707–723. doi:10.1016/S0309-1708(01)00005-7

Lawless, C., M.A. Semenov, and P.D. Jamieson. 2008. Quantifying the effect of uncertainty in soil moisture characteristics on plant growth using a crop simulation model. Field Crops Res. 106:138–147. doi:10.1016/j.fcr.2007.11.004

Lawrence, N.D. 2004. Gaussian process latent variable models for visualisation of high dimensional data. Adv. Neural Inf. Process. Syst. 16:329–336.

Lieberman, C., K. Willcox, and O. Ghattas. 2010. Parameter and state model reduction for large-scale statistical inverse problems. SIAM J. Sci. Comput. 32:2523. doi:10.1137/090775622

Liu, Y.N., M.Y. Hussaini, and G. Okten. 2013. Optimization of a Monte Carlo variance reduction method based on sensitivity derivatives. Appl. Numer. Math. 72:160–171. doi:10.1016/j.apnum.2013.06.005

Liu, Y.N., E. Jimenez, M.Y. Hussaini, G. Okten, and S. Goodrick. 2015. Parametric uncertainty quantification in the Rothermel model with randomised quasi-Monte Carlo methods. Int. J. Wildland Fire 24:307–316. doi:10.1071/WF13097

Lucia, D.J., P.S. Beran, and W.A. Silva. 2004. Reduced-order modeling: New approaches for computational physics. Prog. Aerosp. Sci. 40:51–117. doi:10.1016/j.paerosci.2003.12.001

Marrel, A., B. Iooss, M. Jullien, B. Laurent, and E. Volkova. 2011. Global sensitivity analysis for models with spatially dependent outputs. Environmetrics 22:383–397. doi:10.1002/env.1071

Mascaro, G., E.R. Vivoni, and R. Deidda. 2010. Downscaling soil moisture in the southern Great Plains through a calibrated multifractal model for land surface modeling applications. Water Resour. Res. 46. doi:10.1029/2009WR008855

Maxwell, R.M., F.K. Chow, and S.J. Kollet. 2007. The groundwater–land-surface–atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations. Adv. Water Resour. 30:2447–2466. doi:10.1016/j.advwatres.2007.05.018

McFadden, J.P., F.S. Chapin, and D.Y. Hollinger. 1998. Subgrid-scale variability in the surface energy balance of arctic tundra. J. Geophys. Res., D, Atmospheres 103(D22):28947–28961. doi:10.1029/98JD02400

McGuire, A.D., J.S. Clein, J.M. Melillo, D.W. Kicklighter, R.A. Meier, C.J. Vorosmarty, and M.C. Serreze. 2000. Modelling carbon responses of tundra ecosystems to historical and projected climate: Sensitivity of pan-Arctic carbon storage to temporal and spatial variation in climate. Glob. Change Biol. 6:141–159. doi:10.1046/j.1365-2486.2000.06017.x

Micchelli, C.A., and M. Pontil. 2005. On learning vector-valued functions. Neural Comput. 17:177–204. doi:10.1162/0899766052530802

Montaldo, N., and J.D. Albertson. 2003. Temporal dynamics of soil moisture variability: 2. Implications for land surface models. Water Resour. Res. 39(10). doi:10.1029/2002WR001618

Oberbauer, S.F., J.D. Tenhunen, and J.F. Reynolds. 1991. Environmental effects on CO_2 efflux from water track and tussock tundra in arctic Alaska, U.S.A. Arct. Alp. Res. 23:162–169. doi:10.2307/1551380

Oechel, W.C., S.J. Hastings, G. Vourlitis, M. Jenkins, G. Riechers, and N. Grulke. 1993. Recent change of arctic tundra ecosystems from a net carbon-dioxide sink to a source. Nature 361:520–523. doi:10.1038/361520a0

Oleson, K.W., D.M. Lawrence, G.B. Bonan, B. Drewniak, M. Huang, C.D. Koven, S. Levis, F. Li, W.J. Riley, Z.M. Subin, S.C. Swenson, P.E. Thornton, A. Bozbiyik, R. Fisher, E. Kluzek, J.-F. Lamarque, P.J. Lawrence, L.R. Leung, W. Lipscomb, S. Muszala, D.M. Ricciuto, W. Sacks, Y. Sun, J. Tang, and Z.-L. Yang. 2013. Technical description of version 4.5 of the Community Land Model (CLM). NCAR Technical Note. National Center for Atmospheric Research, Boulder, CO.

Olson, R., R. Sriver, M. Goes, N.M. Urban, H.D. Matthews, M. Haran, and K. Keller. 2012. A climate sensitivity estimate using Bayesian fusion of instrumental observations and an Earth System model. J. Geophys. Res. 117(D4):D04103. doi:10.1029/2011JD016620

Paniconi, C., and M. Putti. 1994. A comparison of Picard and Newton iteration in the numerical-solution of multidimensional variably saturated flow problems. Water Resour. Res. 30:3357–3374. doi:10.1029/94WR02046

Pau, G.S.H., G. Bisht, and W.J. Riley. 2014. A reduced-order modeling approach to represent subgrid-scale hydrological dynamics for land-surface simulations: Application in a polygonal tundra landscape. Geosci. Model Dev. 7:2091–2105. doi:10.5194/gmd-7-2091-2014

Pau, G.S.H., Y. Zhang, and S. Finsterle. 2013. Reduced order models for many-query subsurface flow applications. Computat. Geosci. 17:705–721. doi:10.1007/s10596-013-9349-z

Prud'homme, C., D.V. Rovas, K. Veroy, L. Machiels, Y. Maday, A.T. Patera, and G. Turinici. 2002. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. J. Fluids Eng. 124(1):70–80 (Trans. ASME). doi:10.1115/1.1448332

Quarteroni, A., G. Rozza, and A. Manzoni. 2011. Certified reduced basis approximation for parametrized partial differential equations and applications. J. Math. Industry 1(1):3. doi:10.1186/2190-5983-1-3

Rasmussen, C.E., and H. Nickisch. 2010. Gaussian processes for machine learning (GPML) Toolbox. J. Mach. Learn. Res. 11:3011–3015.

Rasmussen, C.E., and C.K.I. Williams. 2006. Gaussian processes for machine learning. MIT Press, Cambridge, MA.

Ratto, M., A. Castelletti, and A. Pagano. 2012. Emulation techniques for the reduction and sensitivity analysis of complex environmental models. Environ. Model. Softw. 34(C):1–4. doi:10.1016/j.envsoft.2011.11.003

Razavi, S., B.A. Tolson, and D.H. Burn. 2012a. Numerical assessment of metamodelling strategies in computationally intensive optimization. Environ. Model. Softw. 34:67–86. doi:10.1016/j.envsoft.2011.09.010

Razavi, S., B.A. Tolson, and D.H. Burn. 2012b. Review of surrogate modeling in water resources. Water Resour. Res. 48. doi:10.1029/2011WR011527

Rewienski, M., and J. White. 2003. A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. IEEE Trans. Comput. Aided Des. Integrated Circ. Syst. 22:155–170. doi:10.1109/TCAD.2002.806601

Richards, L.A. 1931. Capillary conduction of liquids through porous mediums. J. Appl. Phys. 1(1):318–333.

Riley, W.J., and C. Shen. 2014. Characterizing coarse-resolution watershed soil moisture heterogeneity using fine-scale simulations. Hydrol. Earth Syst. Sci. 18:2463–2483. doi:10.5194/hess-18-2463-2014

Robinson, D.A., C.S. Campbell, J.W. Hopmans, B.K. Hornbuckle, S.B. Jones, R. Knight, F. Ogden, J. Selker, and O. Wendroth. 2008. Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review. Vadose Zone J. 7:358–389. doi:10.2136/vzj2007.0143

Robinson, T., M. Eldred, K. Willcox, and R. Haimes. 2012. Strategies for multifidelity optimization with variable dimensional hierarchical models. 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. American Institute of Aeronautics and Astronautics, Reston, VA.

Rodriguez-Iturbe, I., A. Porporato, L. Ridolfi, V. Isham, and D.R. Cox. 1999. Probabilistic modelling of water balance at a point: The role of climate, soil and vegetation. Proc. R. Soc. Math. Phys. Eng. Sci. 455:3789–3805. doi:10.1098/rspa.1999.0477

Rodriguez-Iturbe, I., G.K. Vogel, R. Rigon, D. Entekhabi, F. Castelli, and A. Rinaldo. 1995. On the Spatial-Organization of Soil-Moisture Fields. Geophys. Res. Lett. 22:2757–2760. doi:10.1029/95GL02779

Rougier, J., D.M.H. Sexton, J.M. Murphy, and D. Stainforth. 2009. Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. J. Clim. 22:3540–3557. doi:10.1175/2008JCLI2533.1

Rowley, C.W., T. Colonius, and R.M. Murray. 2004. Model reduction for compressible flows using POD and Galerkin projection. Physica D 189:115–129. doi:10.1016/j.physd.2003.03.001

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. Design and analysis of computer experiments. Statistical science. A review. J. Inst. Math. Stat. 4:409–435.

Samaniego, L., R. Kumar, and M. Zink. 2013. Implications of parameter uncertainty on soil moisture drought analysis in Germany. J. Hydrometeorol. 14:47–68. doi:10.1175/JHM-D-12-075.1

Sanderson, B.M., R. Knutti, T. Aina, C. Christensen, N. Faull, D.J. Frame, W.J. Ingram, C. Piani, D.A. Stainforth, D.A. Stone, and M.R. Allen. 2008. Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. J. Clim. 21:2384–2400. doi:10.1175/2008JCLI1869.1

Saridakis, K.M., and A.J. Dentsoras. 2008. Soft computing in engineering design—A review. Adv. Eng. Inform. 22:202–221. doi:10.1016/j.aei.2007.10.001

Schilders, W.H.A., H.A.d. Vorst, and J. Rommes. 2008. Model order reduction theory, research aspects and applications. Mathematics in Industry 13. The European Consortium for Mathematics in Industry. Springer, Berlin.

Schuur, E.A.G., J. Bockheim, J.G. Canadell, E. Euskirchen, C.B. Field, S.V. Goryachkin, S. Hagemann, P. Kuhry, P.M. Lafleur, H. Lee, G. Mazhitova, F.E. Nelson, A. Rinke, V.E. Romanovsky, N. Shiklomanov, C. Tarnocai, S. Venevsky, J.G. Vogel, and S.A. Zimov. 2008. Vulnerability of permafrost carbon to climate change: Implications for the global carbon cycle. Bioscience 58:701–714. doi:10.1641/B580807

Simpson, T., J. Peplinski, P. Koch, and J. Allen. 2001. Metamodels for computer-based engineering design: Survey and recommendations. Eng. Comput. 17:129–150. doi:10.1007/PL00007198

Teh, Y.W., M. Seeger, and M.I. Jordan. 2005. Semiparametric latent factor models. In: R. Cowell and Z. Ghahramani, editors, Proceedings of Tenth International Workshop on Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics, NJ. p. 333–340.

Teuling, A.J., and P.A. Troch. 2005. Improved understanding of soil moisture variability dynamics. Geophys. Res. Lett. 32:L05404.

Western, A.W., R.B. Grayson, and G. Bloschl. 2002. Scaling of soil moisture: A hydrologic perspective. Annu. Rev. Earth Planet. Sci. 30:149–180. doi:10.1146/annurev.earth.30.091201.140434

Wilkinson, R.D. 2010. Bayesian calibration of expensive multivariate computer experiments In: L. Biegler et al., editors, Large-scale inverse problems and quantification of uncertainty. John Wiley & Sons, New York. p. 195–215.

Willcox, K., and J. Peraire. 2002. Balanced model reduction via the proper orthogonal decomposition. AIAA J. 40:2323–2330. doi:10.2514/2.1570

Zona, D., D.A. Lipson, R.C. Zulueta, S.F. Oberbauer, and W.C. Oechel. 2011. Microtopographic controls on ecosystem functioning in the Arctic Coastal Plain. J. Geophys. Res. Biogeosci. 116. doi:10.1029/2009JG001241