SIAM/ASA J. Uncertainty Quantification
Vol. 2, pp. 545–563

Model Calibration via Deformation*

William Kleiber†, Stephan R. Sain‡, and Michael J. Wiltberger§

Abstract. Dynamical computer models often exhibit space-time features that are partially misaligned or misshapen when compared to observational data. Whether due to approximate numerical schemes, incomplete physics, or estimated boundary conditions, the goal of calibrating these models to field data involves optimally aligning model output with observed features. The traditional approach to correcting model discrepancy is to introduce an additive and/or multiplicative bias. Especially for dynamical models, systematic bias may alternatively be viewed as deformation bias. We introduce an expanded approach to model calibration in the presence of space-time feature discrepancy. Borrowing ideas from the image warping literature, we propose a nonlinear deformation of the computer model that optimally aligns with observed images; probabilistically this manifests as a transformation of model coordinate space with a variational penalty on the likelihood function. We apply the approach to a dynamical magnetosphere-ionosphere computer model that exhibits challenging feature discrepancies, and we successfully identify a region of input parameter space that simultaneously minimizes model error and discrepancy from field data.

Key words. calibration, computer experiment, deformation, discrepancy, image warping, misalignment, space-time displacement

AMS subject classifications. 62M30, 65D18, 97K70

DOI. 10.1137/130935367

1. Introduction.

Dynamical space-time computer models have become a fundamental staple of science and society, ranging from numerical weather prediction, climate science, hydrological, and pollution forecasting to the ecological and environmental sciences. These computer models are built by solving nonlinear physical equations that define model evolution. Often solutions are unable to exactly capture the physical phenomena of interest, and so tuning parameters to approximate unknown physics or account for incomplete numerical schemes are included. A common characteristic of these models is that particular input settings will generate spatial features that partially replicate observational features, whereas other input settings may not exhibit the observational features at all. Let us motivate this problem with the following example.

Consider the set of observational data and model output from a computer model shown in Figure 1. A typical question posed by a geophysical scientist is, Which example model output, (a), (b), or (c) in Figure 1, best represents the observational data? Heuristically, it appears that the output of panel (a) does not adequately reproduce the magnitude of the

*Received by the editors September 3, 2013; accepted for publication (in revised form) July 11, 2014; published electronically September 23, 2014.
http://www.siam.org/journals/juq/2/93536.html

†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309 (william.kleiber@colorado.edu).
‡Geophysical Statistics Project, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO 80307 (ssain@ucar.edu).
§High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO 80307 (wiltbemj@ucar.edu).

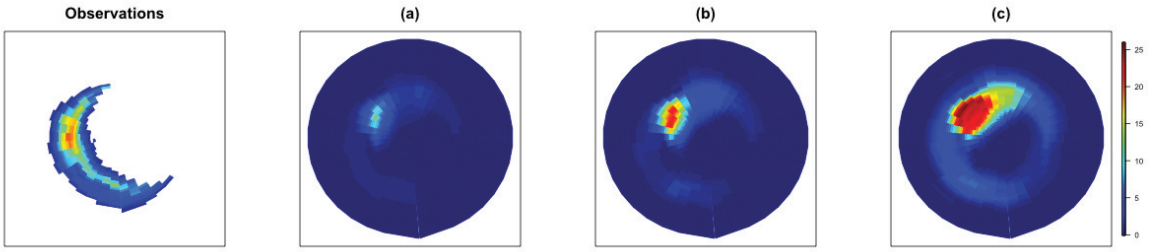

Figure 1. *Observations and model output at three different input settings. Units are keV. A typical scientific question is, Which version of model output,* (a), (b), *or* (c), *best represents the observational data? Displayed observations are satellite derived data of a geomagnetic storm in the near-Earth space environment, with the LFM-MIX as the comparing physical model; see the body of the paper for specific discussion.*

observational data. Panel (b) may have approximately the correct magnitude of the mode but exhibits some spatial displacement and does not adequately capture the spatial extent of the feature. The output in (c) exhibits exaggerated magnitude that is spatially displaced from the observations. On the other hand, panel (c) may exhibit a closer-to-correct spatial feature, albeit a spatially smoothed and displaced version thereof. This type of geographical discrepancy is common in geophysical models, especially numerical weather prediction models, space weather models, and regional or global climate models [9]. Note that the exact dataset in use in Figure 1 compares satellite observations of a geomagnetic storm occurring in Earth's near-space environment, and the physical model is the LFM-MIX, which we discuss in greater detail below in section 4.

An additional complication to the question posed by Figure 1 is that all displayed model output is from the same deterministic geophysical model but at three different sets of parameter tunings. A common goal is to find the "best" set of input parameters such that the output optimally aligns with observational data, i.e., to calibrate the computer model. Our goal in this paper is to explore a framework for calibration in the scenario that the computer model exhibits displacement and/or deformation discrepancy as compared to field data. A scientist may examine Figure 1 and decide, based on expert understanding, that the output of panel (b) is most accurate, as it has approximately the same magnitude of the mode but also requires a rotation and deformation correction. While this may be sufficient for exploratory purposes, a more complete statistical solution will quantify these notions and the uncertainty in such an analysis.

The motivation for our approach follows from the field of forecast verification in the atmospheric sciences [9]. To validate (i.e., quantify the quality of) the three forecasts shown in Figure 1, we value the panels such that the geographical features of the forecast field and corresponding observational field are similar in both geographical location and spatial pattern. To quantify "similar," deformation approaches are a focus of recent interest [16]. The basic idea is to geographically deform the model output of panels (a), (b), and (c) to match the observations (in both spatial pattern and location), with some quantification of the "work" or "energy" required to match the images. Then, panels that require less "energy" to align the two images are considered to be superior forecasts. For example, in Figure 1, to spatially align

the model output and observational data, we might expect a greater effort to align panels (a) and (c), which both require rotations as well as substantial nonlinear deformations that will still have difficulty identifying the field magnitude, whereas the output of panel (b) visually appears to more easily align with the observations with a rotation and nonlinear stretching about latitude bands and is of similar magnitude with the observations.

The idea of deformation is quite different from the traditional approach to modeling discrepancy between model/forecast fields and observational fields, which in the statistical literature is to include (potentially space-time varying) additive and/or multiplicative discrepancy processes [6, 25, 32]. While additive and multiplicative processes are flexible modeling tools, they do not readily quantify adequate replication of spatial features as directly as a deformation approach. Our approach generalizes the traditional statistical approach, putting greater emphasis on model tunings that properly replicate observed spatial patterns, incorporating notions of geographical deformation from the image warping literature in the context of model calibration.

The method of calibration we introduce relies on spatially deforming the model output so that relevant features spatially match those in the field data. Our technique relies on graphical features of computer models and lies at the intersection of computer experiments and image warping [2]. The field of image warping or deformation has roots in multiple disciplines for a range of scientific aims, including image registration [26], forecast verification [16, 9], shape analysis [18, 22], edge detection [12], and spatial statistics [34, 1, 8]. In the next section we cover a brief introduction to the basic idea of image warping and set up some useful notation. Following this, we present a statistical model for computer experiments, discussing estimation and calibration. The subsequent section contains an application of our approach to a space weather geophysical model that exhibits challenging spatial displacement discrepancy, and the paper concludes with suggestions for future research.

2. Image warping.

The main idea in this paper is to introduce some tools from the image warping literature for a computer experiments goal, using deformation to correct and quantify model bias. Before delving into calibration and modeling details, we cover some introductory notions from image warping and set up notation that will be useful in the following uncertainty quantification development.

For this section, suppose the practitioner has two images, say the observations and image (b) from Figure 1, and the scientific objective is to deform, or warp, panel (b) so that it most closely aligns with the observations. This heuristic idea may be mathematically viewed as redefining the geographical indexing of panel (b) in such a way that the magnitudes of each pixel more closely match those of the observations. In effect, deformations typically involve image rotations, stretches, and squeezes to achieve alignment.

Denote by $Y(s)$ the observed image in Figure 1 and by $M(s)$ the model image of panel (b) for $s = (x, y)' \in \mathcal{D} \subset \mathbb{R}^2$. For exposition, we deal with planar images in this section, although later we suggest some extensions to warping in higher dimensions.

The observations $Y(s)$ are represented as a noisy version of a properly deformed model image,

$$Y(s) = M\big(W(s)\big) + \varepsilon(s),$$

where $\varepsilon(s)$ is a white noise process, and $W : \mathbb{R}^2 \to \mathbb{R}^2$ is the warping or deformation function. The effect of W is to reindex the image so that each pixel's magnitude more closely matches that of the observed image Y, noting that W consists of two separate deformations, one in the x-coordinate and one in the y-coordinate.

In the image warping literature, it is widely acknowledged that estimating a deformation function such as W without any constraints can lead to nonphysical warps, such as space and/or time folding over upon itself [7, 17]. To guard against this possibility, a penalty inner product is typically introduced, $\langle \cdot, \cdot \rangle$, on a reproducing kernel Hilbert space of functions, $\mathcal{H} = \{f | f : \mathbb{R}^2 \to \mathbb{R}^2, \langle f, f \rangle < \infty\}$. A special example is a penalty that results in thin plate splines [35]. Procrustes transformations, useful for shape analysis, act by allowing rotations and magnification of the image [18]. Additionally, there are penalties yielding diffeomorphism solutions [13].

If observations are available at n locations s_i, $i = 1, \ldots, n$, then the warping function is traditionally estimated by minimizing

$$\frac{1}{n\tau^2} \sum_{i=1}^{n} \Big(Y(s_i) - M\big(W(s_i)\big) \Big)^2 + \lambda \langle W(\cdot), W(\cdot) \rangle, \tag{2.1}$$

where $\lambda \in \mathbb{R}^+$ is a tuning or regularization parameter and τ^2 is a variance parameter (noting that operationally τ^2 is a nuisance parameter).

As noted above, the term $\langle W(\cdot), W(\cdot) \rangle$ can be viewed as a penalty on possible warping functions, and is chosen to preserve desirable properties of the warp, such as continuity or differentiability [7, 34]. Inflated values of λ imply greater penalties, and thus result in estimated warps that exhibit less deformation, while small values of λ allow for more substantial deformation to accurately align M to Y. An alternative statistical interpretation of $\langle \cdot, \cdot \rangle$ is that it defines a (log)-prior on the space of warping functions, and thus (2.1) has an equivalent and attractive interpretation as a negative log-likelihood [17].

In the image warping literature, solutions minimizing (2.1) are constrained to a representation involving a linear combination of the kernel function of the Hilbert space $\mathcal{H}$ [3, 4, 14]. In our example, with reproducing kernel $R(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}^+$, the jth component of the vector function $W = (W_1, W_2)' = (W_x, W_y)'$ is

$$W_j(s) = \sum_{i=1}^{m} a_{ij} N_i(s) + \sum_{i=1}^{\ell} b_{ij} R(s, s_{0i}), \tag{2.2}$$

where $\{N_i\}_{i=1}^{m}$ span the null space of $\mathcal{H}$ [35]. Prototypical examples of null space functions are rigid rotations and shifts when the penalty involves the Laplacian. The locations $s_{01}, \ldots, s_{0\ell}$ are so-called landmarks in the domain and are usually chosen by the practitioner to align with important features on the image. Naturally, $s_{0i} = s_i$ is a possibility, using a set of basis functions centered at all image locations, but in practice this introduces a large number of statistical parameters b_{ij}, and landmarks are chosen on a case-by-case basis.

The particular difficulty in our context is that, for instance, in Figure 1, each image arises from the same geophysical model and is generated via different parameter tunings for the model. Below we explore the natural extension of deforming single images by allowing the

deformation function W to be dependent on these parameter tunings. This extension sets up an extended framework for calibrating computer models exhibiting systematic deformation bias.

3. Statistical model. Consider a space-time computer model that is also indexed by a set of input parameters, $\theta \in \mathcal{D}$, where the space in which θ exists, $\mathcal{D}$, is usually a subset of $\mathbb{R}^k$ or $\mathbb{Z}^k$ for some k or some direct product thereof. The θ parameter vector often includes physical constants or tuning variables that must be chosen for the code to run. While expert knowledge can sometimes inform best values of the parameters, there is often great uncertainty regarding the optimal values, and a statistical approach to tuning these parameters to field data is often necessary and desired. We write the computer model output as $M(z, \theta)$ for space-time location $z = (s', t)' \in \mathbb{R}^d \times \mathbb{R}$, where naturally $s \in \mathbb{R}^d$ is spatial location and $t \in \mathbb{R}$ denotes time point. Suppose the physical process of interest has corresponding observations $Y(z)$ at space-time locations $z = z_1, \ldots, z_n$.

A fundamental problem in this context is to assess potential model bias. In particular, even at some "best" setting θ_0, it is possible and typically understood that the computer model M is not a perfect representation of reality. In other words, a model such as

$$Y(z) = M(z, \theta_0) + \varepsilon(z), \tag{3.1}$$

where $\varepsilon(z)$ is a white noise measurement error process, is not an appropriate statistical model. To account for potential model bias, various authors have suggested modeling frameworks that are encompassed by the following generic model,

$$Y(z) = \beta_0(z) + \beta_1(z)M(z, \theta_0) + \varepsilon(z), \tag{3.2}$$

where $\beta_0(z)$ and $\beta_1(z)$ are additive and multiplicative model biases, respectively [6, 25, 32] (as before, $\varepsilon(z)$ is often modeled as a white noise measurement error process). While this approach is highly flexible, the bias processes β_0 and β_1 do not directly incorporate possible geographical biases such as location bias or deformation discrepancy; our proposal in the next section generalizes the approach of (3.2).

3.1. Space-time deformation model. Whereas computer models may exhibit fixed additive and multiplicative biases, in the space-time context it is often more crucial to view nonstationary model bias as a space-time discrepancy. Heuristically, the model may exhibit spatial features similar to observed features but possibly appearing deformed or warped as compared to observations. Thus a careful measurement of goodness-of-fit of the computer model should reward the fact that the space-time features are present in the model output, but it should also reward the input settings at which these features are closest in geographical space to the field data. To make the idea firm, we propose the following statistical model:

$$Y(z) = \beta_0(z) + \beta_1(z)M(W(z, \theta_0), \theta_0) + \varepsilon(z). \tag{3.3}$$

Contrasting (3.2) with (3.3), the newly introduced function $W : \mathbb{R}^{d+1} \times \mathcal{D} \to \mathbb{R}^{d+1}$ is a space-time deformation function. The inclusion of W acknowledges that the computer model may exhibit spatial and/or temporal displacement of important observational features, but

we still allow for potential remaining additive and multiplicative discrepancy. This framework allows for the possibility of model bias but also includes the case when the computer model is unbiased. In the case of no bias, $\beta_0(z) = 0$, $\beta_1(z) = 1$, and at the best setting θ_0 the warping function is the identity when restricted to $\mathbb{R}^{d+1}$, so $W(z, \theta_0) = z$.

In many space-time applications where often only one realization of the space-time process is available, the proposed model (3.3) is overparameterized. Teasing out the difference between space-time varying additive/multiplicative biases and deformations will typically require expert knowledge regarding the error structure of the computer experiment at hand. A direct simplification that results in identifiability is to set the linear biases to be space-time constant, so

$$Y(z) = \beta_0 + \beta_1 M(W(z, \theta_0), \theta_0) + \varepsilon(z). \tag{3.4}$$

We adopt this modification of the model through the remainder of the statistical development, but we acknowledge that the fully flexible (3.3) is available given sufficient field data to distinguish between the two main types of model bias.

Estimation of the warping function follows the same approach as is typical for individual image warping. For a fixed value of θ, the optimal warping function $W(\cdot, \theta) \in \mathcal{H}$ is the function that minimizes

$$G(\theta) = \frac{1}{n\tau^2} \sum_{i=1}^{n} \Big(Y(z_i) - \big(\beta_0 + \beta_1 M(W(z_i, \theta), \theta)\big) \Big)^2 + \lambda \langle W(\cdot, \theta), W(\cdot, \theta) \rangle, \tag{3.5}$$

where $\lambda \in \mathbb{R}^+$ is a tuning or regularization parameter and τ^2 is a variance parameter (as previously, τ^2 is a nuisance parameter). Setting λ to be large emphasizes smooth warping functions, while a small value allows for greater warping in order to reduce the error of warped model output. In the case of no model bias, at the unknown setting θ_0, W is the identity map, and the criterion (3.5) reduces to the sum of squared errors of model residuals. If W is nontrivial, we say M is biased.

The second term of (3.5) is a crucial element in our approach, not only for regularizing the estimation process but also for interpreting calibration results. The penalty can be interpreted as a quantification of the amount of work, or bending energy, required to correct the deformation bias of the computer model [7]. Together, the two terms making up G combine to emphasize parameters exhibiting little-to-no deformation bias and such that, after having accounted for this bias, the model minimizes the sum of squared residuals from the field data.

3.1.1. Deformation function. Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and define by $R(\cdot, \cdot) : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$ the associated reproducing kernel. The jth component of the vector function $W = (W_1, \ldots, W_{d+1})'$ is

$$W_j(z, \theta) = \sum_{i=1}^{m} a_{ij}(\theta) N_i(z) + \sum_{i=1}^{\ell} b_{ij}(\theta) R(z, z_{0i}), \tag{3.6}$$

where $\{N_i\}_{i=1}^m$ span the null space of $\mathcal{H}$ [35]. The locations $z_{01}, \ldots, z_{0\ell}$ are the landmarks, and as noted previously, $\ell = n$ is a possibility, however, and for purposes of calibration it will be useful to opt for a more parsimonious model.

Momentarily suppressing the temporal dynamics, the most common spatial dimension is the plane, $d = 2$, so $z = s = (s_x, s_y)' \in \mathbb{R}^2$. In this dimension, a common penalty is that which results in a thin plate spline solution [19, 35],

$$\langle f, f \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\left(\frac{\partial^2 f}{\partial s_x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial s_x \partial s_y} \right)^2 + \left(\frac{\partial^2 f}{\partial s_y^2} \right)^2 \right) \mathrm{d}s_x \mathrm{d}s_y,$$

on the space of twice-differentiable functions where $\langle f, f \rangle < \infty$. This inner product has the desirable interpretation of a bending energy [7]. In this case, the two components of the warping function $W = (W_x, W_y)'$ can be written as in (3.6):

$$W_j(s, \theta) = a_{1j}(\theta) + a_{2j}(\theta) s_x + a_{3j}(\theta) s_y + \sum_{i=1}^{\ell} b_{ij}(\theta) R(s, s_{0i})$$

for $j = x, y$, where the radial basis function is $R(s, s_{0i}) = \|s - s_{0i}\|^2 \log\left(\|s - s_{0i}\|^2\right)$, $\|\cdot\|$ denoting the usual Euclidean distance function. We can choose any set of ℓ landmark points, and the mapping of these determines the movement of all other locations in the domain, assuming known coefficients $a_{ij}(\theta)$ and $b_{ij}(\theta)$. While $d = 2$ is the most common spatial dimension, some spatial computer models may be additionally indexed by elevation or higher dimensions; in these cases the thin plate spline formulation extends directly to $\mathbb{R}^d$; see [35] for relevant discussion.

Operationally, the thin plate spline yields a useful algebraic framework for deformation. Define an $\ell \times \ell$ matrix $K = (R(s_{0i}, s_{0j}))_{i,j=1}^{\ell}$, an $\ell \times 3$ matrix P whose ith row is $(1, s_{0ix}, s_{0iy})$, and

$$L = \begin{pmatrix} K & P \\ P' & O \end{pmatrix}, \tag{3.7}$$

where O is a 3×3 matrix of zeroes. We require the following conditions for $j = x, y$, ensuring that W_j has square integrable derivatives [35]:

$$\sum_{i=1}^{\ell} b_{ij}(\theta) = \sum_{i=1}^{\ell} b_{ij}(\theta) s_{0ix} = \sum_{i=1}^{\ell} b_{ij}(\theta) s_{0iy} = 0.$$

Thus, given any set of warped coordinates $W(s_i, \theta) = (W_x(s_i, \theta), W_y(s_i, \theta))'$, we can succinctly write the procedure as

$$L \begin{pmatrix} b_{1j}(\theta) \\ \vdots \\ b_{\ell j}(\theta) \\ a_{1j}(\theta) \\ a_{2j}(\theta) \\ a_{3j}(\theta) \end{pmatrix} = \begin{pmatrix} W_j(s_1, \theta) \\ \vdots \\ W_j(s_\ell, \theta) \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.8}$$

for $j = x, y$. For the thin plate spline formulation, the penalty term, or bending energy, can be written as $W_j(\theta)' L_0 K L_0 W_j(\theta)$ where L_0 is the upper left $\ell \times \ell$ submatrix of L^{-1}, and $W_j(\theta) = (W_j(s_{01}, \theta), \ldots, W_j(s_{0\ell}, \theta))'$.

Two technical difficulties arise in a formulation such as this. The first occurs when a warped coordinate $(W_x(s, \theta), W_y(s, \theta))'$ does not exactly align with an original image coordinate; in this case a common approach in the image warping literature is to bilinearly interpolate warped values from the four surrounding grid locations to the warped coordinate (an alternative may be to use a statistical emulator of the computer model). Second, if the warped coordinate lies outside of the domain of interest, we artificially set the warped model output to zero; for the geomagnetic storm example below this would simply imply that the storm occurs only in the available geographical domain, which is a reasonable assumption. For more discussion on the details of image warping and these technical obstacles, the reader is directed to [15].

3.2. Calibration.

We turn to using the above statistical model for calibrating the computer model M. Calibration involves finding the optimal value or set of values of the input parameter vector θ such that the computer model is a best representation of reality. The most straightforward approach to calibration is to directly minimize $G(\theta)$, perhaps using an approach in the same spirit as [31], where fast calibration of computer models by interpolating summary criteria is considered. However, even at the "best" choice of θ_0 corresponding to the minimized G surface, without additional modeling there is no straightforward way to emulate the model or warping function at θ_0, and thus the practitioner is unable to easily check the result. We propose an approach that includes a more complete representation of the various sources of uncertainty, using an emulator for posterior validation.

3.2.1. Calibration via emulation.

To properly account for the uncertainty in predicting the G criteria at a new input setting θ, we must account for the fact that the computer model has not yet been run at this setting, and thus an analysis should quantify the uncertainty in predicting $M(\cdot, \theta)$. A more detailed notation for G is $G(\theta | M, W)$, as the exact value of G is known only when the computer model output and appropriate deformation function at setting θ are known. The first step is to build an emulator, or statistical surrogate, for the computer model M. Numerous approaches have been proposed to this end, and emulation of complex dynamical models is an active area of current research [10, 11, 23, 28, 33].

On top of an emulator, we must also predict an appropriate warping function at an untried input setting. To this end, momentarily restrict attention to only spatial deformations that may depend on time point. We suggest a slight variation on the warping function decomposition (3.6). For $z = (s', t)'$ and $j = x, y$, let

$$W_j(z, \theta) = \sum_{i=1}^{m} a_{ij}(\theta, t) N_i(s) + \sum_{i=1}^{\ell} b_{ij}(\theta, t) R(s, s_{0i}), \tag{3.9}$$

where we now allow the coefficients a_{ij} and b_{ij} to vary with input setting and time. This setup allows the spatial deformation to be a function of input setting but also acknowledges the case where the deformation evolves temporally. Naturally one can consider reversing the roles of location s and time point t in (3.9) if it is most convenient to think of a set of temporal

misalignments that are spatially indexed. A preferable but very difficult alternative is to simultaneously deform space and time with $R(\cdot,\cdot)$ and $N_i(\cdot)$ being functions of z.

The coefficients a_{ij} and b_{ij} are modeled as independent constant mean Gaussian processes. We parameterize the covariance function for each coefficient process, explicitly taking account of temporal correlation, as

$$C(\theta_1,t_1;\theta_2,t_2)=\sigma^2\mathrm{M}_2\left(\frac{|t_1-t_2|}{\xi_t}\right)\prod_{i=1}^{p}\mathrm{M}_2\left(\frac{|\theta_{1i}-\theta_{2i}|}{\xi_i}\right), \tag{3.10}$$

where M_2 is a Matérn correlation function with smoothness 2, σ^2 is a variance, and ξ_i is a correlation scale parameter corresponding to the ith component input parameter [20]. The Matérn correlation function continuously indexes the smoothness of the associated random field; in the limiting case that the smoothness is infinite, the Matérn coincides with the squared exponential correlation. The smoothness is notoriously difficult to estimate, and using a fixed value of 2 corresponds to the assumption that coefficients are continuously differentiable across input settings. Each instance of a_{ij} and b_{ij} has distinct spatial parameters of mean and $\sigma,\xi_t,\xi_1,\ldots,\xi_p$. To estimate the spatial parameters, we minimize (3.5) for each θ_p, $p=1,\ldots,19$, yielding estimates $\hat{a}_{ij}$, $\hat{b}_{ij}$. Conditional on these point estimates, the spatial parameters are estimated by maximum likelihood, exploiting the Gaussianity assumption.

We are now set up to describe the calibration procedure. Suppose an emulator has been chosen, $\hat{M}(z,\theta)$, that is either stochastic or deterministic. We set up a Metropolis–Hastings algorithm for sampling from the penalized log-likelihood $-G(\theta|M,W)$, with normal proposal densities centered at the current parameter setting. For simplicity, and as is often the case in scientific studies, we impose a uniform prior on θ over some hyperrectangle of scientifically plausible values. Then, at a new proposal value, θ_n, emulate $\hat{M}(z,\theta_n)$ either deterministically or by sampling. Likewise, predict the appropriate warp $\hat{W}(\cdot,\theta_n)$ via a conditional expectation using the Gaussian process model for the coefficients in (3.6). Calibration is completed by performing the Metropolis–Hastings step on the new penalized log-likelihood value $-G(\theta_n|\hat{M},\hat{W})$ compared to the current value at θ_{n-1}.

The computational challenge of our modeling framework is diminished by a judicious choice of model as well as estimation approach. For instance, general estimation of parameters of a deformation function such as (3.6) can be challenging; we suggest using a parsimonious parameterization with as few landmarks ℓ as possible such that the deformation function still yields sufficient flexibility for the experiment at hand (below we find five landmark points to be sufficient for our scientific example). Additionally, estimation of the spatial covariance parameters in (3.10) can be a computational burden. The number of covariance parameters that must be estimated is tempered by choosing a small number of landmarks. In our experience, fixing the smoothness at a reasonable value and compartmentalizing estimation of the remaining spatial covariance parameters to a separate step, as opposed to simultaneous estimation within a fully Bayesian framework, helps reduce overall computational burden. Indeed, as the deformation coefficients are assumed to be independent, it is straightforward to parallelize estimation of the corresponding covariance parameters.

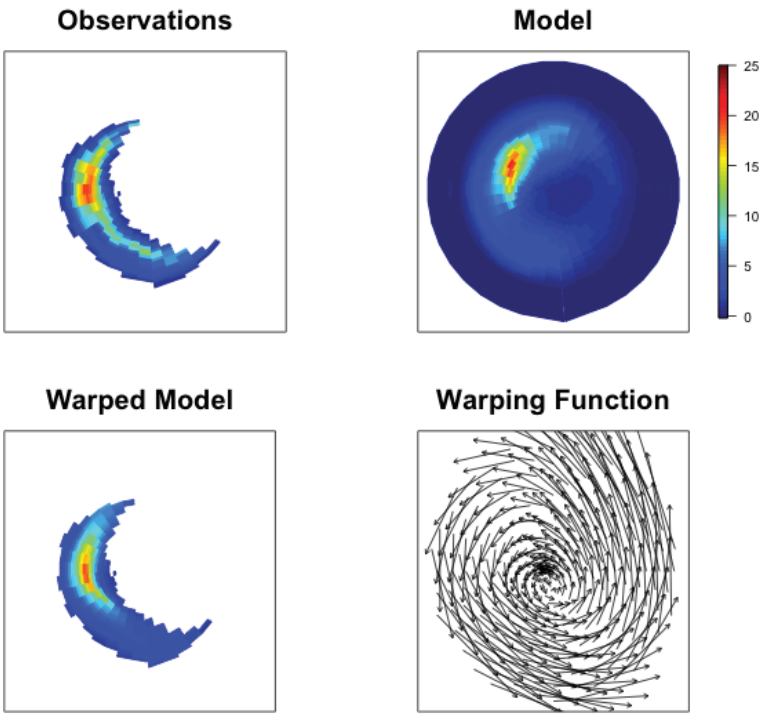

Figure 2. *Observations, LFM-MIX model output, warped model output, and corresponding warping function. Units are keV.*

4. Calibrating a space weather model. Geomagnetic storms can have surprisingly drastic impacts on modern society, affecting electrical systems, flight routing, and GPS accuracy. A leading geophysical model for geomagnetic storm modeling (among other space weather facets) is the Lyon–Fedder–Mobarry model [29]. Linked with the ionospheric MIX model [30], the LFM-MIX computer model is a state-of-the-art modeling system for the near-Earth space environment.

The LFM-MIX is parameterized by three tuning parameters, $\theta = (\alpha, \beta, R)$, whose exact values are unknown and require estimation or calibration by tuning model output to field data [36]. The observational dataset we work with derives from satellite measurements of a geomagnetic storm occurring on January 10, 1997 over the North Pole. Observations $Y(z)$ are a space-time field at 18 equally spaced time points over a two hour period on a spatial grid of between 261 and 314 points, the exact number of spatial observations depending on time point.

The LFM-MIX is a substantial testbed for the proposed methodology, exhibiting spatial displacement bias that evolves across time and depends on input setting. Figures 2 and 3 illustrate the complex space-time structures of the LFM-MIX and corresponding observations at two different time points. Indeed the clear spatial discrepancy in both figures is difficult to accommodate using available techniques in the computer experiments literature.

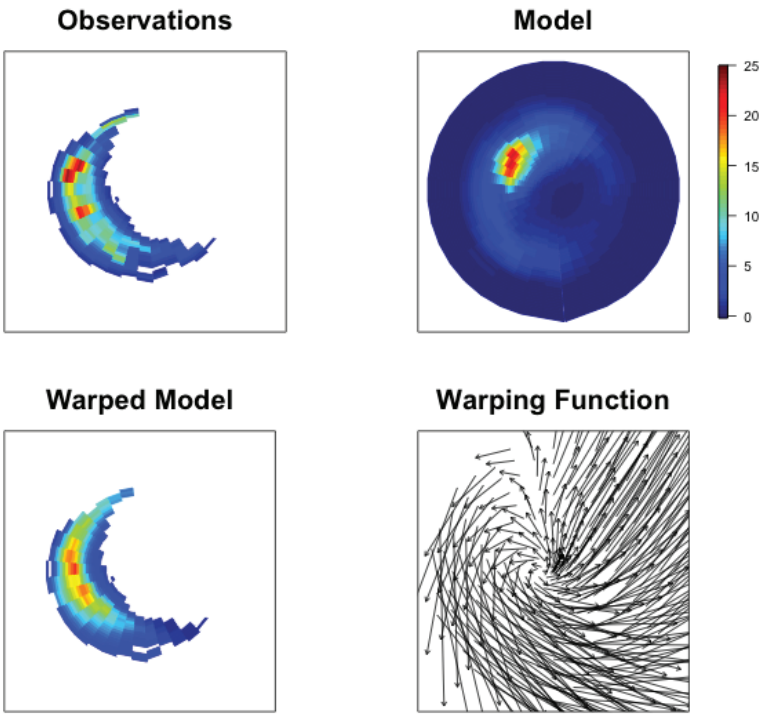

Figure 3. *Same as Figure 2 but for a different time point.*

4.1. Experimental design. The LFM-MIX is a computationally intensive model. It is available at multiple resolutions, of which we use the coarsest for computational feasibility. The coarsest runs with a wall-clock time corresponding approximately to real time. Accounting for model spin-up, each separate run of the LFM-MIX requires approximately 18 hours to implement. We use a space-filled design of runs of the LFM-MIX at 20 points in (α, β, R) space over the domain $[0, 0.5] \times [0, 2.5] \times [0, 0.1]$, defining physically meaningful intervals for the three parameters [31]. We remove one run at the lower boundary from consideration, as expert opinion finds the model output to be unphysical; this leaves us with $p = 19$ initial design points.

4.2. Statistical model and emulator. We begin by entertaining the same observation model as (3.4), but with $\beta_0 = 0$, as physically it is inappropriate to allow the additive bias to be nonzero. The class of warping functions we opt for are the thin plate spline functions of section 3.1.1. Exploratory analysis suggests there is no substantial evidence of temporal displacement bias, whereas the spatial displacement is clear; see Figures 2 and 3. Hence, the particular class of basis functions is spatial only, with $W(z, \theta) = (W_x(s, \theta), W_y(s, \theta))'$.

To perform any spatial deformation, we first require landmark points. Note that we could choose landmarks to be spatially constant across each image, or alternatively we may choose image-specific landmarks that correspond to the same spatial features (e.g., placing one landmark at the mode of each image which may be in a different spatial location per

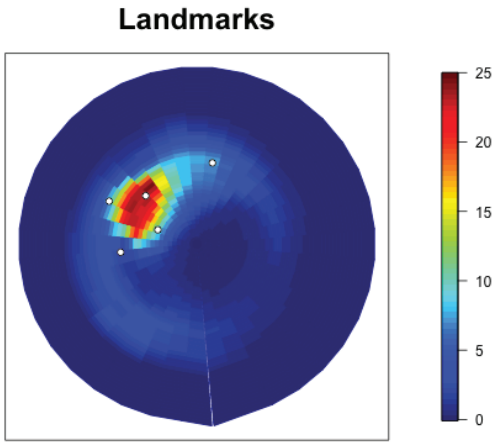

Figure 4. *Landmark points shown as five white dots.*

image). The danger in the latter approach is the possibility that not all images share the same spatial features, where the choice of which spatial features to focus on becomes unclear. For the LFM-MIX, we choose five landmarks from the first time point of the run $(\alpha, \beta, R) = (0.23, 1.39, 0.052)$, shown in Figure 4. This particular run is chosen as the landmark template, as it is centrally located (in θ space) amongst all 19 runs. The five displayed landmarks were chosen as a group of four outlining the approximate boundary of the high energy spatial feature that is part of the auroral zone that has physical significance for the model, with the fifth landmark being placed approximately at the spatial mode. The number of landmark points is ideally chosen to be enough so that interesting and effective deformations are possible, but few enough to maintain a parsimonious statistical model, as within the framework of our model each additional landmark introduces numerous statistical parameters.

Although the warps we consider are only in the spatial dimension, we allow for temporal evolution of the deformation function and use the coefficient model of (3.9). The coefficients a_{ij} and b_{ij} are modeled as constant mean Gaussian processes. Each instance of a_{ij} and b_{ij} is endowed with a covariance of the form (3.10) and has distinct spatial parameters $\sigma, \xi_\alpha, \xi_\beta, \xi_R, \xi_t$. Minimizing the negative penalized likelihood (3.5) for each $\theta_p, p = 1, \ldots, 19$ yields estimates $\hat{a}_{ij}, \hat{b}_{ij}$. Conditional on these point estimates, we estimate the spatial parameters by maximum likelihood.

To perform calibration, we require an emulator for the computer model. The LFM-MIX produces high dimensional space-time model output, with 29160 space-time locations per input setting. With our initial design this results in 554040 total points of model output, and hence a dimension reduction approach is desired. We follow the emulator of [27], which was also developed for the LFM-MIX, an approach we briefly describe here. The basic idea is to decompose the LFM-MIX output as weighted sums of spatial empirical orthogonal functions (EOFs). The EOFs are the left principal component vectors of the matrix whose columns are composed of vectorized spatial fields of model output for a single time point and input

setting. In particular, we statistically model $M(z,\theta)$, for $z=(s',t)'$, as

$$M(z,\theta)=\sum_{i=1}^{3} v_i(t,\theta)U_i(s) \tag{4.1}$$

using the first three spatial EOFs $U_i(s)$. The first three EOFs capture more than 99% of the variability of the LFM-MIX. We model the EOF weights as Gaussian processes. The first EOF coefficient process is endowed with a mean function

$$\mathbb{E}v_1(t,\theta)=\gamma_0+\gamma_1\alpha+\gamma_2 R\sqrt{\alpha}+\gamma_3\cos(2\pi t/18)+\gamma_4\sin(2\pi t/18). \tag{4.2}$$

The harmonics are included to account for the natural rise and fall in magnitude of the geomagnetic storm over the two hour storm period. The linear and nonlinear functions of input parameters are informed by model physics. The remaining two EOFs are assumed to be mean zero. Finally, each EOF coefficient is assumed to have a covariance of the form (3.10), where naturally each of the three EOF coefficients has distinct parameters. Briefly, mean function parameters are estimated by ordinary least squares, and conditional on these the spatial parameters of variance and ranges are estimated by maximum likelihood. The interested reader can find a detailed description of how the model physics and EOF structure correspond to likelihood assumptions in [27]. Note that similar modeling strategies have been employed in [23, 24].

4.3. Results. We adopt the statistical model of section 4.2 for the LFM-MIX and explore model fit as well as calibration results. The first step is to estimate statistical parameters. Estimating warping function parameters $\{a_{ij}(\theta_p)\}_{i,j}$ and $\{b_{ij}(\theta_p)\}_{i,j}$ can be computationally challenging, especially if all model runs $p=1,\ldots,19$ are estimated simultaneously. We adopt an estimation scheme that retains computational feasibility via a series of conditional stages. For the following exposition, we fix the regularization parameter λ at 1, which approximately equally weights the penalty and average squared error between warped model and observations, based on exploratory analyses. Results also are under the assumption that bias manifests only as deformation bias, where the additive and multiplicative biases $\beta_0=0$ and $\beta_1=1$ are based on exploratory analysis; see below for further discussion.

The first step is to marginally estimate warping function parameters $\{a_{ij}(\theta_p)\}_{i=1}^{m}$ and $\{b_{ij}(\theta_p)\}_{i=1}^{\ell}$ for $j=x,y$ by penalized maximum likelihood, minimizing $G(\theta_p)$ for $p=1,\ldots,19$. This step is readily parallelized, and thus for experiments with larger initial designs a potential bottleneck can be avoided. Conditional on these estimates, the spatial parameters are estimated by maximum likelihood. The emulator is estimated as discussed in section 4.2, leaving only the input parameter (α,β,R) to estimate. Recall that the version of the LFM-MIX in use is a coarse resolution, and hence we expect the physical model to reproduce large scale space-time features of the observed geomagnetic storm but caution that fine scale structures are not expected to be captured in the current implementation.

Before turning to calibration, let us examine the behavior of the deformation component. Figures 2 and 3 include warped model output with corresponding deformation functions that optimally align with the observations for two different time points. Clearly the optimal deformation function evolves across time, generally requiring a clockwise rotation along with

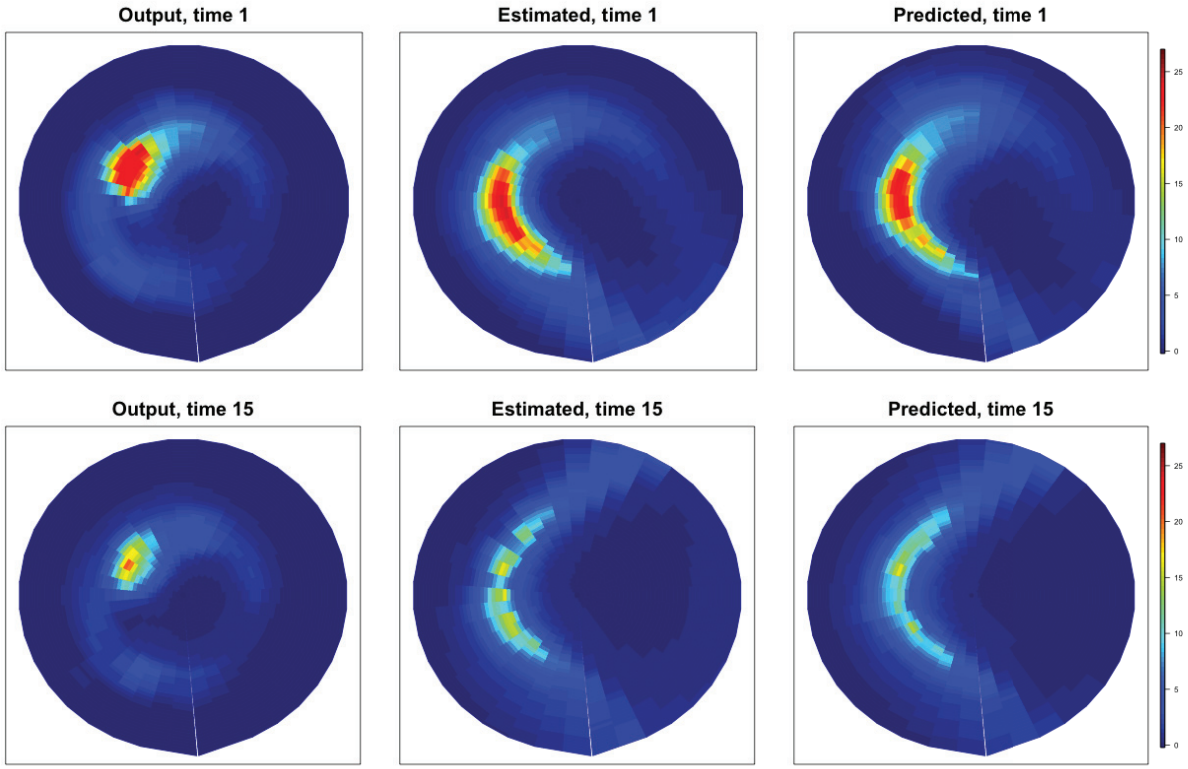

Figure 5. *Model output $M(z,\theta)$ (first column), optimal warped function estimated directly from the model output $M(W(z,\theta),\theta)$ (second column), and predicted model output with predicted deformation function $\hat{M}(\hat{W}(z,\theta),\theta)$ (third column) for example time points* 1 *and* 15. *Units are keV.*

nonlinear squeezing and stretching. Note that even after deformation the model may not be a perfect representation of the field data; within the statistical model the remaining discrepancy is modeled as a white noise process.

The stochastic model for the deformation function W of (3.9) allows us to predict appropriate warping functions for untried input settings θ. To examine our ability to predict feasible deformations for a new setting, we note that two predictions are required—that for the warping function W, and that for the computer model itself, i.e., the emulator. Figure 5 illustrates the ability of our construction to adequately interpolate in this complex scenario. The first column contains two sets of model output at example time points 1 and 15, while the second column is the estimated warp based on these known sets of model output, i.e., $M(W(z,\theta),\theta)$. The final column contains the emulated model and predicted deformation corresponding to these two time points, $\hat{M}(\hat{W}(z,\theta),\theta)$, via a leave-one-run out prediction process, with $\hat{M}$ and $\hat{W}$ representing conditional means based on the respective Gaussian process representations. Visually, the predicted warping function accurately captures the salient features of the optimal warp—the clockwise rotation and nonlinear stretching present for time point 1, and a similar rotation and slightly more pronounced nonlinear stretch at time point 15. Also notice the emulated model is an accurate representation of the LFM-MIX, reproducing the correct

Table 1

RMSE values for quantities based on experimental design of 19 input settings. First row is leave-one-out cross-validation comparison of emulator to true LFM-MIX solution ($M(z,\theta) - \hat{M}(z,\theta)$). Remaining rows contain predictive RMSE of the LFM-MIX and the emulator compared to the observations averaged over all storm time points; second and third rows are no deformation of LFM-MIX/emulator; and fourth row is RMSE for warped emulator.

	1	2	3	4	5	6	7	8	9	10
LFM-MIX − Emulator	0.64	0.45	0.39	0.35	0.15	0.34	0.09	0.34	0.56	0.45
Obs − LFM-MIX, no warp	7.41	6.38	5.66	6.23	5.95	6.23	5.00	5.69	6.96	6.29
Obs − Emulator, no warp	7.41	6.32	5.55	6.13	5.94	6.07	5.00	5.58	6.87	6.23
Obs − Emulator, warped	7.59	5.13	4.34	4.53	4.10	2.75	4.95	2.94	6.67	4.56
	11	12	13	14	15	16	17	18	19	
LFM-MIX − Emulator	0.22	0.48	0.13	0.40	0.76	0.32	0.86	0.75	0.22	
Obs − LFM-MIX, no warp	3.86	6.58	5.28	6.06	7.24	4.60	8.06	7.74	4.43	
Obs − Emulator, no warp	3.83	6.44	5.28	5.90	7.21	4.55	8.03	7.70	4.40	
Obs − Emulator, warped	2.86	4.14	4.11	3.26	7.82	2.63	9.06	7.86	3.15	

spatial features as well as the magnitude of the energy peak at both time points.

4.3.1. Validating the statistical model. We perform a cross-validation study of the predictive performance of the statistical model. For each of the 19 input settings we hold out individual model runs and compare the emulated model behavior to the true LFM-MIX behavior in terms of predictive root mean squared error (RMSE); Table 1 provides these values. The top row quantifies the error of the emulator based on predictive leave-one-out RMSE. The average error of the emulator from the LFM-MIX is less than 1 keV for almost all runs, averaging at 0.39 keV. Notice the magnitude of this RMSE of emulator prediction is an order of magnitude less than the warped model from observations, so we expect the error in predicting model output to have a negligible effect on calibration. Warping yields substantial improvements on model fit, as the warped model output improves RMSE for 15 of the 19 initial design runs, at an average of 23.5% reduction in RMSE. The cases where the warp inflates the RMSE are characterized by multiple time points where model energy levels are substantially different from the field data, and hence while the spatial features are optimally aligned, the apparent model error is inflated.

4.3.2. Calibration. With all statistical model parameters estimated, we are set up to tune the input vector θ to the observational data. We run three Metropolis–Hastings samplers initialized at random input settings, all of which were found to converge on the same area of parameter space. Independent normal proposal densities centered at the current values of α, β, and R are used. For each chain, we discard a burn-in of 500 samples, and posterior density estimates are shown in Figure 6 with small black dots indicating the initial design points. The posterior densities are relatively concentrated around the mode $(\alpha, \beta, R) = (0.22, 0.75, 0.029)$.

To assess the scientific quality of the posterior distribution, we show an emulated and bias corrected version of model output based on an example sample from the posterior distribution, as well as posterior surfaces of emulated and warped output at the $1\%, 50\%$, and 99% quantiles in Figure 7. The sampled setting of the model exhibits approximately the same magnitude of energy as in the observations but clearly still exhibits some deformation bias.

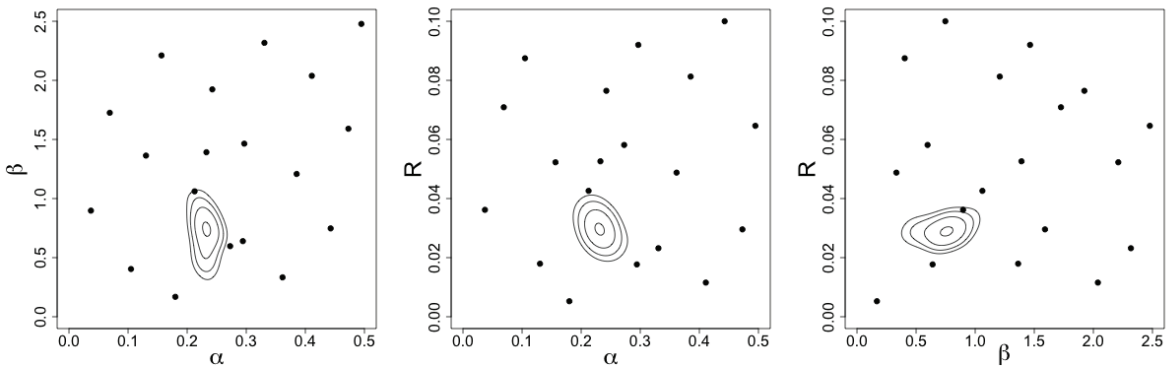

Figure 6. *Pairwise posterior densities with initial design shown as solid black dots.*

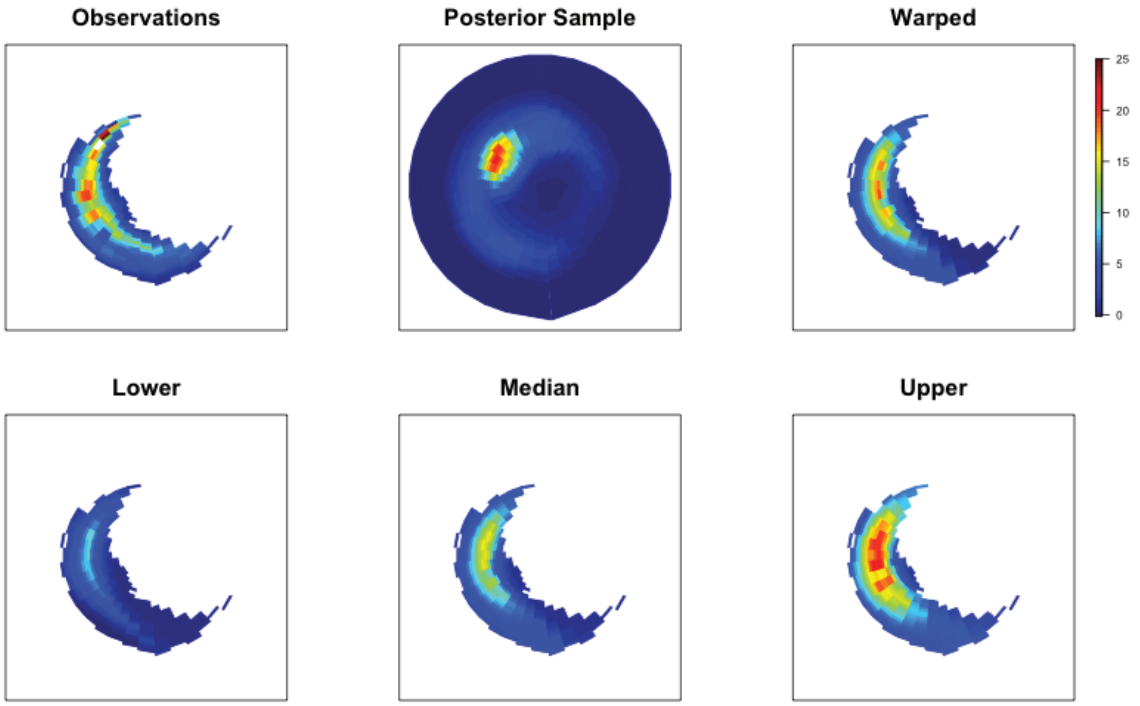

Figure 7. *Observations and pointwise posterior* 1%, 50%, *and* 99% *quantiles of predicted and warped model output. Units are keV.*

The spatially deformed version of the optimal setting matches the spatial features apparent in the observations, accurately capturing the arc of energy and approximate borders of the geomagnetic storm. We remind the reader that the particular version of the LFM-MIX in use is a relatively coarse model, and we do not expect this resolution at any input setting to replicate fine scale structure of the geomagnetic storm, but on the other hand we expect this resolution to represent large scale behavior, albeit a spatially biased version thereof.

The choice of considering only deformation bias and setting the weighting parameter $\lambda = 1$ to approximately balance warped model accuracy and deformation energy is largely based on exploratory analysis. Choosing $\beta_0 = 0$ is justified physically, as the energy in the ionosphere must be approximately zero outside of the storm area. We experimented with allowing β_1 to vary and found posterior calibration distributions that were multimodal and favored no particular region of parameter space. This suggests there is a confounding between the input parameters and potential multiplicative bias, which is expected given that α and R contribute directly and nonlinearly to the magnitude of the storm's energy (see [27] for discussion). We additionally experimented with inflating λ to 10, and the posterior calibration densities updated slightly from those shown in Figure 6, with the posterior mode generally centering around $(0.22, 0.07, 0.027)$. The values of α and R are robustly estimated against this inflation in the deformation energy penalty, but the value of β is shrunk; this suggests that α and R primarily control the magnitude of the geomagnetic storm, while β has influence over the extent of the spatial feature.

5. Discussion. This paper introduces an approach to calibration in the context of dynamical space-time computer models that exhibit displacement bias. The approach is rather different from the traditional statistical technique of accounting for model bias and additive and multiplicative factors, which can also be included within our framework. The basic idea is to consider a transformation of the coordinate space of the model output, which can deform space, time, or both. Optimal warping functions are estimated based on a penalized likelihood formulation at all available initial design points and can be parallelized for easy implementation with large initial designs. We parameterize the warp in such a way that deformation functions can be predicted at new design points without requiring a new run of the computer model. Given an emulator, calibration proceeds as optimizing a penalized log-likelihood.

The LFM-MIX has been previously examined by a number of other authors, including those of this current manuscript. [27] proceeded under the assumption that a higher resolution version of the LFM-MIX is unbiased at an optimal input setting. By linking the lower resolution version (that which is used in this paper) to a higher resolution version (that is computationally more expensive), they were able to identify three regions of parameter space which may contain optima and developed a sequential design based on expected improvement to further hone in on the optimum. Our approach herein does not use the higher resolution model and is able to identify a single region of parameter space that optimally aligns warped low resolution LFM-MIX with observational data. Additionally, our posterior densities approximately overlap with those derived by [27].

A second set of authors from the same scientific working group has developed a predictive process approach to emulation and calibration [5, 21]. The authors again used the higher resolution version of the LFM-MIX and statistically linked it with the current lower resolution version, but they added the assumption that the model, even at optimal input settings, is biased. The bias correction is a spatially and temporally varying additive bias and also includes a rotational bias of the LFM-MIX to better align with field data. The additive bias accounts for high frequency spatial features in the observational data that are not well replicated by the LFM-MIX even at a higher resolution. Our approach here can be thought of as a partial generalization of [21] in that we allow rotations as well as nonlinear deformations

of the model output to better align with the satellite observations.

This paper opens a number of potential future research routes. An initial idea is to develop methodological approaches for linking multiple resolutions of the computer model (à la [27, 21]) but under the assumption of model deformation bias. A second route is to consider warping in the presence of multiple geomagnetic storms; whether the appropriate warps should be considered constant across storms or dependent on the particular dataset in use is an open question. Third, the LFM-MIX actually generates a bivariate space-time field, where herein we focus on the energy variable. Bivariate deformation would be required for a full analysis. Another potential route of future research is optimal siting of landmark locations that not only capture scientifically meaningful spatial features but also lend themselves to statistical modeling in a calibration and interpolation context.

REFERENCES

[1] S. Aberg, F. Lindgren, A. Malmberg, J. Holst, and U. Holst, *An image warping approach to spatio-temporal modelling*, Environmetrics, 16 (2005), pp. 833–848.

[2] S. Allassonnière, Y. Amit, and A. Trouvé, *Towards a coherent statistical framework for dense deformable template estimation*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 69 (2007), pp. 3–29.

[3] Y. Amit, *A nonlinear variational problem for image matching*, SIAM J. Sci. Comput., 15 (1994), pp. 207–224.

[4] N. Arad and D. Reisfield, *Image warping using few anchor points and radial functions*, Comput. Graph. Forum, 14 (1995), pp. 35–46.

[5] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang, *Gaussian predictive process models for large spatial data sets*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 70 (2008), pp. 825–848.

[6] M. J. Bayarri, J. O. Berger, R. Paulo, and J. Sacks, *A framework for validation of computer models*, Technometrics, 49 (2007), pp. 138–154.

[7] F. L. Bookstein, *Principal warps: Thin-plate splines and the decomposition of deformations*, IEEE Trans. Pattern Anal. Mach. Intell., 11 (1989), pp. 567–585.

[8] L. Bornn, G. Shaddick, and J. V. Zidek, *Modeling nonstationary processes through dimension expansion*, J. Amer. Statist. Assoc., 107 (2012), pp. 281–289.

[9] B. G. Brown, E. Gilleland, and E. E. Ebert, *Forecasts of spatial fields*, in Forecast Verification: A Practitioner's Guide in Atmospheric Science, I. T. Jolliffe and D. B. Stephenson, eds., John Wiley & Sons, New York, 2012, pp. 95–117.

[10] S. Conti, J. P. Gosling, J. E. Oakley, and A. O'Hagan, *Gaussian process emulation of dynamic computer codes*, Biometrika, 96 (2009), pp. 663–676.

[11] S. Conti and A. O'Hagan, *Bayesian emulation of complex multi-output and dynamic computer models*, J. Statist. Plann. Inference, 140 (2010), pp. 640–651.

[12] T. F. Cootes and C. F. Taylor, *On representing edge structure for model matching*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2001, pp. 1114–1119.

[13] P. Dupuis, U. Grenander, and M. I. Miller, *Variational problems on flows of diffeomorphisms for image matching*, Quart. Appl. Math., 56 (1998), pp. 587–600.

[14] M. Fornefett, K. Rohr, and H. S. Steihl, *Radial basis functions with compact support for elastic registration of medical images*, Image. Vis. Comput., 19 (2001), pp. 87–96.

[15] E. Gilleland, L. Chen, M. DePersio, G. Do, K. Eilertson, Y. Jin, E. L. Kang, F. Lindgren, J. Lindström, R. L. Smith, and C. Xia, *Spatial Forecast Verification: Image Warping*, Technical Report TN-482+STR, National Center for Atmospheric Research, Boulder, CO, 2011.

[16] E. Gilleland, J. Linström, and F. Lindgren, *Analyzing the image warp forecast verification method on precipitation fields from the ICP*, Weather Forecast., 25 (2010), pp. 1249–1262.

[17] C. A. Glasbey and K. V. Mardia, *A penalized likelihood approach to image warping*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 63 (2001), pp. 465–514.

[18] C. Goodall, *Procrustes methods in the statistical analysis of shape*, J. Roy. Statist. Soc. Ser. B, 53 (1991), pp. 285–339.
[19] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall/CRC, Boca Raton, FL, 1994.
[20] P. Guttorp and T. Gneiting, *Studies in the history of probability and statistics* XLIX: *On the Matérn correlation family*, Biometrika, 93 (2006), pp. 989–995.
[21] M. J. Heaton, W. Kleiber, S. R. Sain, and M. Wiltberger, *Emulating and calibrating the multiple-fidelity Lyon-Fedder-Mobarry magnetosphere-ionosphere coupled computer model*, J. Roy. Stat. Soc. C Appl. Stat., 2014, DOI: 10.1111/rssc.12064.
[22] T. Heimann and H. Meinzer, *Statistical shape models for 3D medical image segmentation: A review*, Med. Image Anal., 13 (2009), pp. 543–563.
[23] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer model calibration using high-dimensional output*, J. Amer. Statist. Assoc., 103 (2008), pp. 570–583.
[24] M. B. Hooten, W. B. Leeds, J. Fiechter, and C. K. Wikle, *Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models*, J. Agr. Biol. Environ. Stat., 16 (2011), pp. 475–494.
[25] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models (with discussion)*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 63 (2001), pp. 425–464.
[26] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, *Deformation models for image recognition*, IEEE Trans. Pattern Anal. Mach. Intell., 29 (2007), pp. 1422–1435.
[27] W. Kleiber, S. Sain, M. Heaton, M. Wiltberger, S. Reese, and D. Bingham, *Parameter tuning for a multi-fidelity dynamical model of the magnetosphere*, Ann. Appl. Stat., 7 (2013), pp. 1286–1310.
[28] F. Liu and M. West, *A dynamic modelling strategy for Bayesian computer model emulation*, Bayesian Anal., 4 (2009), pp. 393–412.
[29] J. G. Lyon, J. A. Fedder, and C. M. Mobarry, *The Lyon-Fedder-Mobarry (LFM) global MHD magnetospheric simulation code*, J. Atmos. Sol. Terr. Phys., 66 (2004), pp. 1333–1350.
[30] V. G. Merkin and J. G. Lyon, *Effects of the low-latitude ionospheric boundary condition on the global magnetosphere*, J. Geophys. Res., 115 (2010), DOI: 10.1029/2010JA015461.
[31] M. T. Pratola, S. R. Sain, D. Bingham, M. Wiltberger, and J. Rigler, *Fast sequential model calibration of large nonstationary spatial-temporal processes*, Technometrics, 55 (2013), pp. 232–242.
[32] P. Z. G. Qian and C. F. J. Wu, *Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments*, Technometrics, 50 (2008), pp. 192–204.
[33] J. Rougier, *Efficient emulators for multivariate deterministic functions*, J. Comput. Graph. Stat., 17 (2008), pp. 827–843.
[34] P. D. Sampson and P. Guttorp, *Nonparametric estimation of nonstationary spatial covariance structure*, J. Amer. Statist. Assoc., 87 (1992), pp. 108–119.
[35] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
[36] M. Wiltberger, R. S. Weigel, W. Lotko, and J. A. Fedder, *Modeling seasonal variations of auroral particle precipitation in a global-scale magnetosphere-ionosphere simulation*, J. Geophys. Res., 114 (2009), DOI: 10.1029/2008JA013108.