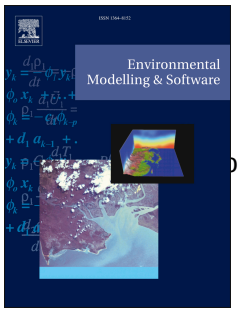
Journal Pre-proof



Developing short-term probabilistic fo of meso-temperature fields

Byeongseong Choi, Mario Berges, Elie B Matteo

PII: S1364-8152(21)00231-0

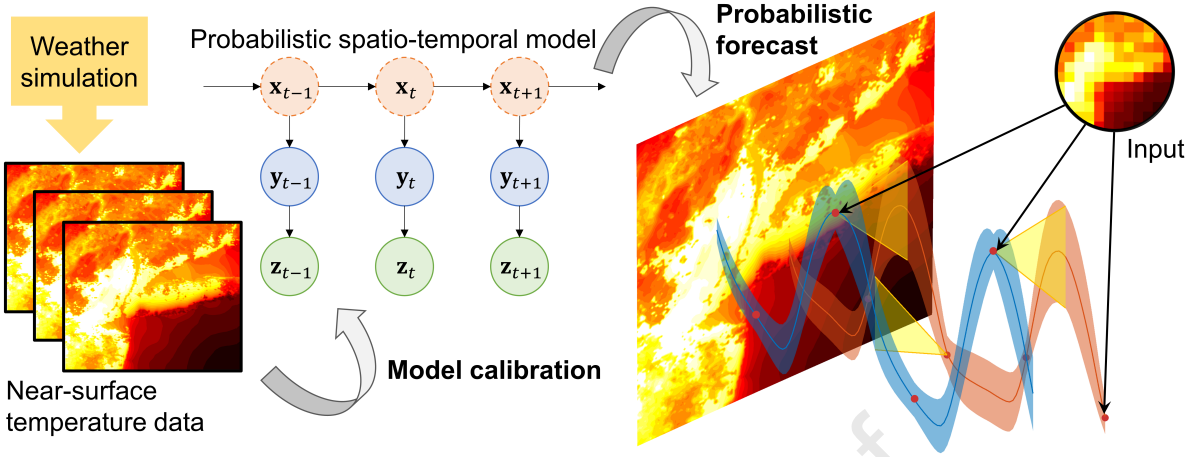
DOI: https://doi.org/10.1016/j.envsoft.2021.105189

Reference: ENSO 105189

To appear: *Environmental Modelling and Software*

Accepted: 2 September 2021

Please cite this article as: Choi, B., Berges, M., Bou probabilistic forecasts of meso-scale ensemble numerical weather fields, Environmental Modelling and Software, https://doi.org/10.1016/j.envsoft.2021.105189



Highlights

Developing short-term probabilistic forecasts of meso-scale near-surface urban temperature fields

Byeongseong Choi,Mario Berges,Elie Bou-Zeid,Matteo Pozzi

- We develop probabilistic models for meso-scale near-surface urban air temperature.
- We calibrate/validate the models on simulated data for New York City and Pittsburgh.
- A Kalman filter/smoother updates the proposed models for adaptive forecast.
- The proposed models use 3 to 8% the computing resources used by a comparable model.
- 24-hours ahead forecasts show 0.97-1.13°C prediction error.

Developing short-term probabilistic forecasts of meso-scale near-surface urban temperature fields

Byeongseong Choi[a,*], Mario Berges[a], Elie Bou-Zeid[b] and Matteo Pozzi[a,*]

[a]Department of Civil and Environmental Engineering, Carnegie Mellon University, Porter Hall 119, Pittsburgh, 15213-3890, PA, U.S.A.
[b]Department of Civil and Environmental Engineering, Princeton University, E414 Engineering Quadrangle, Princeton, 08544, NJ, USA

ARTICLE INFO

Keywords:
urban heat
probabilistic model
spatio-temporal model
latent space
state-space model
Kalman filter

ABSTRACT

This paper introduces a probabilistic approach to spatio-temporal high resolution meso-scale modeling of near-surface temperature in regions of dimension about 150km∼200km, with 1km grid spacing and 30-minutes interval. Such probabilistic models can accurately forecast short-term temperature fields and serve as a computationally less expensive alternative to physics-based models that necessitate high-performance computing. The probabilistic models here are calibrated from simulations of a physics-based model, the Princeton Urban Canopy Model, coupled to the Weather Research and Forecasting Model (WRF-PUCM). We assess the performance of the calibrated models to forecast short-term near-surface temperature in various cases. In the numerical campaign, our models achieve 0.97-1.13°C root mean squared error (RMSE) for 24 hours ahead forecast; generating three days of forecast takes between 20 and 170 seconds on a single processor computer. Hence, the proposed approach provides predictions at relatively high accuracy and low computational cost.

In the document below we report in blue the changes made to the article since its last submission.

1. Introduction

Temperature has a remarkable influence on human society ranging from heat-related deaths to economic loss. Heatwaves are reportedly one of the deadliest weather-related hazards in the United States [6, 25] and killed more than 160,000 people around the globe between 1998 and 2017, including an annual spike of 72,000 deaths in Europe in 2003 [33]. The economic impact of extreme temperature includes extra energy consumption, increased water use, lost productivity, and other additional costs in the built environment [27, 13, 38]. Due to the high population and vulnerability to heat stressors, urban areas have more relevant applications for high-resolution surface temperature forecasts. By trapping heat within the building canopy and through other physical mechanisms, such as reduced evapotranspiration, urban areas increase their near-surface temperature[1] higher than that of their surrounding environment (an effect known as Urban Heat Island, or UHI). UHI has been shown to have a non-linear synergy with extreme heatwaves such that an urban area can be exposed to more severe heat stressors under certain circumstances [17].

Numerical weather prediction (NWP) is the golden-standard for weather research and forecast, both in research and practice [36]. NWP is based on a numerical model that solves partial differential equations for atmospheric fluid dynamics, radiation transfer, cloud process, and other physical processes. With the growth of computing power, NWP has improved its accuracy and, as a result, highly reliable weather forecasts have become available routinely. The Weather Research and Forecasting Model (WRF) is a representative approach for NWP used by the National Weather Service and researchers worldwide [30]. For the broad application of the system, numerous unresolved physics models

*Corresponding author
byeongsc@andrew.cmu.edu, phone: +1-412-417-2832 (B. Choi); mberges@cmu.edu (M. Berges); ebouzeid@princeton.edu (E. Bou-Zeid); mpozzi@cmu.edu, phone: +1-412-268-5649, fax: +1-412-268-7813 (M. Pozzi)
https://www.marioberges.com/ (M. Berges); http://efm.princeton.edu/ (E. Bou-Zeid); https://faculty.ce.cmu.edu/pozzi/ (M. Pozzi)
ORCID(s): 0000-0003-0777-6904 (B. Choi); 0000-0003-2948-9236 (M. Berges); 0000-0002-6137-8109 (E. Bou-Zeid); 0000-0002-9727-2824 (M. Pozzi)

[1]Hereafter, the term "near-surface temperature" refers to the 2m air temperature, which measures ambient air temperature at 2m above the ground. We select this quantity because it has more direct influences on human comfort and mortality than the strictly defined "surface temperature" [1, 32].

have been developed and coupled to WRF. For example, Princeton urban canopy model (PUCM) modifies the default single-layer urban canopy model by considering heterogeneity within urban facets, and urban hydrology [34]. The numerical scheme, integrating WRF and PUCM (WRF-PUCM), has shown high prediction performance, with a bias as small as 0.4∼1.1°C [17, 34, 18, 24]. Still, the main challenge of NWP is its significant computational cost preventing it from being widely used for real-time forecasts or analysis at the city scale, which requires high spatial (and/or temporal) resolution (e.g., for a grid spacing in the order of 1km).

To extend NWP, there have been attempts to improve the spatial resolution of coarse-grid simulation using statistical analysis, called Statistical Downscaling (SD) methods, instead of running additional fine-grid simulations. SD methods estimate local climate variables using statistical relationships with large-scale predictors [22]. SD covers a wide range of models, including regression-based models, bias correction methods, weather-pattern-based classification (weather typing), and stochastic weather generating models [22, 37]. However, to our knowledge, regression-based models, which discover empirical relationships between regional climate variables (predictors) and local ones (predictands), are the most popular method for SD because of their straightforward implementation [23]. According to the discussions in [22, 16], another benefit of SD includes its ability to avoid systematic errors of NWP, embedded in local-scale climate science, especially when it is trained on real weather observations. Many SD methods have been proposed. For example, authors of [26] proposed a downscaling strategy that uses deep convolutional neural networks (CNNs) to attain higher horizontal resolution from NWP output. A method based on support vector machine (SVM) was suggested in [29] to downscale urban air temperature. In spite of their successful outcomes, the methods mentioned above lack a way to increase temporal resolution. As a pure data-driven approach for weather forecast, another CNN-based approach was proposed in [36] to predict one or two fundamental meteorological fields at 500âĂŘhPa geo-potential height, which is widely adopted representative height for the climate status. Even though the method in [36] showed remarkable potentials of deep learning in weather forecast, one still needs a model that enables versatile inference from various types of inputs. For example, a pre-processing technique may be required to the model in [36], as well as SD models, if the input has a missing value or irregular shape, so that a flexible data assimilation is desired. Also, a neural network is a black-box model with intricate structure and parameters, so preference still exists to the comprehensible description on the system.

Probabilistic spatio-temporal models (PSTM) characterize any complex time-evolving process over a spatial domain with probabilistic descriptions. Although both PSTM and SD are based on probability theory, PSTM models a spatio-temporal structure of the process, while most classical SD models learn a mapping from regional climate variables to local ones, as discussed in the previous paragraph. PSTM is flexible in its usage; it can perform inference, even with partially missing inputs. In [4], a Bayesian hierarchical model is proposed to estimate the tropical Pacific sea-surface monthly temperature, capturing the non-linearity in El NiÃśo and La NiÃśa [10]. A Gaussian process (GP) model is developed in [19] to approximate the urban surface temperature, and it is used to optimize the sensing locations over a city, via a Value-of-Information analysis [20]. However, the GP-based approach would need careful applications when high heterogeneity is expected within the field of the climate variables.

In this paper, we introduce a PSTM approach to enable a rapid and efficient forecast of near-surface temperature, with adequate accuracy, as a potential alternative to computationally intensive NWP. The proposed method can capture complex spatial-temporal patterns of temperature with simple representation. Despite mandatory model training, the calibrated model leverages intricate patterns to forecast near-surface temperature from the input measurements, without use of high-performance computing. The probabilistic method is a flexible spatio-temporal downscaling that assimilates diverse types of inputs. The proposed approach is based on linear Markov process with low dimensional latent states, being segmented by the time of the day. In the adopted structure, the latent states are introduced to generate full-scale temperature fields from low dimensional space, so the structure is suitable for faster data assimilation and probabilistic forecast. The model structure also has an advantage in mitigating the risk of overfitting with high dimensional data, e.g., meso-scale near-surface temperature fields, especially when the available data are relatively few.

We develop two alternative linear models. In the first model, the latent state is described by principal components by interpreting them as global features of the temperature field (Principal-component-analysis model; PCA model). The second model selects a set of optimal locations to predict the rest of the temperature field, then the latent state is the temperatures at the selected locations, as the locally focusing feature of the system (Optimal-sensing model; OS model). For each model, we also propose an appropriate calibration method with respect to the given model definition. These two alternative models are suggested to investigate and compare the performance on global features, as in the PCA model, and on local ones, as in the OS model.

We calibrate the models in two meso-scale regions (150km∼200km) around New York City (NYC) and Pittsburgh (PGH) with 1km grid spacing and 30-minutes intervals. After calibrating the spatio-temporal models, we integrate them into a Kalman Filtering/Smoothing scheme to forecast near-surface temperature by processing collected field data and approximated forecasts of external sources. For example, in our numerical test, the proposed models downscale 12km grid spacing data with varying temporal resolutions into 1km spacing temperature fields with 30-minutes intervals. We then compare the accuracy of the models and assess the prediction performance under various case studies.
This paper is an application of the linear Gaussian state-space model to a spatio-temporal downscaling problem, calibrating the model parameters of the hidden Markov model and then performing Kalman filtering/smoothing to forecast short-term near-surface temperature. In the context of this application, our contributions can be summarized as (1) developing techniques for latent variables selection, (2) proposing model calibration methods, and (3) assessing model performance.

2. Probabilistic spatio-temporal model

PSTM needs an efficient structure to capture dependencies among random variables. In this section, we provide a probabilistic description for the process of meso-scale near-surface temperature. Section 2.1 motivates the model structure, and Section 2.2 describes a state-space representation for the spatio-temporal process. Then, Section 2.3 presents two alternative models based on global and on local features, and Section 2.4 provides details on parameter estimation for the proposed models. Finally, Section 2.5 provides background for assimilating data, collected in the field or received from external sources, to forecast near-surface temperature.

2.1. Motivation

The near-surface temperature of an urban region (of dimension about 20km×20km) is influenced by the heat transfer in a much larger domain (of dimension about 200km×200km). The (discretized) temperature field of this large domain is described by a high number of variables. For example, a 150km×150km squared shape domain, discretized with 1km spacing grids, yields $P = 22{,}500$ degrees of freedom in the temperature field for each timestamp. Such high dimensional data pose various technical problems when calibrating a statistical model. For instance, calibrating the full $P \times P$ transition functions requires a tremendous amount of training data, not available in most cases. The calibration needs to rely on a relatively small dataset, so rank deficiency and overfitting are common issues in this problem. To overcome these issues, imposing sparsity in the transition matrix is one of the solutions, and methods have been recently developed to impose sparsity, as in [8]. However, imposing sparsity would increase additional computational complexity, and we are particularly interested in identifying an efficient model structure that reduces the number of model parameters. Thus, in this paper, we propose probabilistic models using a hierarchical architecture, as illustrated in Figure 1.

2.2. State-space model

State-space approaches have been widely adopted as a recursive scheme for estimating the state of a system, and, when the system is linear, they are the base for a variety of Kalman Filter/Smoother schemes [35, 3]. Adopting a state-space description, we model a vector of near-surface temperature field $\mathbf{y}_t$ at time t , of high dimension [P×1], as a linear function of latent variable $\mathbf{x}_t$ of dimension [R×1] with R much lower than P. The vector $\mathbf{y}_t$ is decomposed into three additive terms: an average temperature field $\boldsymbol{\mu}_\tau$, a linear function of the low-dimensional latent states $\mathbf{x}_t$, and an error term $\mathbf{v}_t$:

$$\mathbf{y}_t = \boldsymbol{\mu}_\tau + \boldsymbol{\Phi}\,\mathbf{x}_t + \mathbf{v}_t \tag{1}$$

where $\boldsymbol{\Phi}$ [P×R] is an embedding matrix, and the average temperature field depends on the time of the day τ, discretized into h steps from 00:00 to 24:00, e.g., for $h = 48$, $\tau \in \{00{:}00, 00{:}30, 01{:}00, ..., 23{:}30\}$. τ is defined as $\tau = \mathrm{HoD}(t)$ where $\mathrm{HoD}(\cdot)$ is a function that returns the time of the day of timestamp t. In this hierarchical structure, the dynamics of high dimensional temperature fields is approximated by a linear system in low dimension. The low dimensional representation reduces the number of parameters and the memory requirements in the inference process. Both the latent variables and the error term are modeled as Gaussian random variables. The error term $\mathbf{v}_t$ is assumed as a set of independent and identically distributed random variables with zero mean and standard deviation $\sigma_{\mathbf{v}}$. The temporal

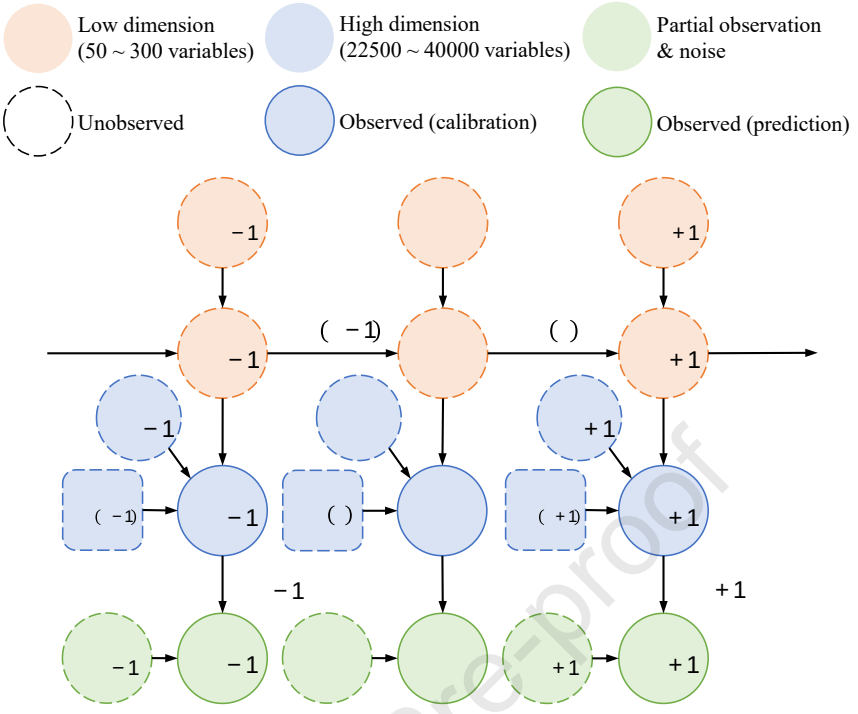


Figure 1: Hierarchical recurrent graph of the proposed model

evolution for the process is assumed to follow a linear Markov model whose transition matrix depends on the time of the day τ.

$$\mathbf{x}_t = \mathbf{F}_\tau \, \mathbf{x}_{t-1} + \mathbf{w}_t \tag{2}$$

where $\mathbf{F}_\tau$ is the transition matrix related to τ, and $\mathbf{w}_t$ is the zero-mean Gaussian noise, related to the transition between times $t-1$ and t, whose covariance matrix $\mathbf{\Sigma}_{\mathbf{w}}^{\tau}$ is also a function of τ.
A measurement vector $\mathbf{z}_t$ is then assumed to be linear combinations of the temperature field $\mathbf{y}_t$.

$$\mathbf{z}_t = \mathbf{H}_t \, \mathbf{y}_t + \boldsymbol{\epsilon}_t \tag{3}$$

where $\boldsymbol{\epsilon}_t$ is an observation noise, often modeled as an zero-mean uncorrelated Gaussian random variable with its covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$. $\mathbf{H}_t$ is an observation matrix. Figure 1 summarizes the overall hierarchical structure of the proposed model.

2.3. Global and local definition of latent state

Given the model structure, there exist infinite solutions to estimate model parameters [3]. The redundancy in solution space can be eliminated by imposing a rule to identify latent variables. Such rule is related to the definition of latent variables, so the model performance depends on what definition is set up for latent variables.
In this paper, we propose two alternative models that employ different latent variables selection schemes. The first model uses principal component analysis (PCA) as a criterion to estimate latent states, focusing on global features of the data as in [10, 14, 31] (PCA model). PCA identifies a set of linearly transformed coordinates (principal components; PCs) of the original data in order of decreasing variance. In this model, PCs are interpreted as global features of the temperature field, so the model uses PCs as the latent states of the process. In addition to the PCA model, we propose an alternative model by defining the state variables as local measures on the temperature field, called the optimal sensing model (OS model). In the OS model, we identify optimally sensing locations which provide the most informative

measures to interpolate/extrapolate the entire temperature field. This model uses such local measures as the latent states of the system. These two alternative models are suggested to guarantee identifiability, then we study which approach (global vs. local features) is a reasonable solution for selecting the latent variables.

2.4. Parameter estimation

In this subsection, we provide procedures to calibrate the model parameters, i.e., $\boldsymbol{\mu}_\tau$, $\boldsymbol{\Phi}$, $\mathbf{F}_\tau$, $\boldsymbol{\Sigma}_{\mathbf{w}}^{\tau}$, for every τ, and $\sigma_{\mathbf{v}}$, from the data. We assume that the proposed models can access the full temperature fields $\mathbf{y}_t$ during the parameter estimation phase. In other words, the model calibration is based on the full temperature fields $\mathbf{y}_t$ instead of the observation $\mathbf{z}_t$, as the variable $\mathbf{z}_t$ is assumed to be a noisy observation on a subset or local average of $\mathbf{y}_t$ with the known observation matrix $\mathbf{H}_t$. We also consider the observation matrix $\mathbf{H}_t$ as a time-varying quantity for the flexible applications.

2.4.1. Mean field

The mean field $\boldsymbol{\mu}_\tau$ can be estimated by simply averaging temperatures at time of the day τ, but such mean fields may contain empirical fluctuation. Therefore, we use a moving average scheme to obtain the smooth empirical mean field $\hat{\boldsymbol{\mu}}_\tau$.

$$\hat{\boldsymbol{\mu}}_\tau = \sum_{\upsilon} \theta_{\tau,\upsilon}\, \bar{\boldsymbol{\mu}}_\upsilon \;\text{ where }\; \bar{\boldsymbol{\mu}}_\upsilon = \frac{1}{N_\upsilon} \sum_{\{t|\mathrm{HoD}(t)=\upsilon\}} \mathbf{y}_t. \tag{4}$$

where N_υ is the number of recorded timestamps with time of the day υ. $\theta_{\tau,\upsilon}$ is a weight coefficient related to the similarity between two times of the day, τ and υ. We adopt a weight that is proportional to an exponentially decaying function with increasing time difference.

$$\theta_{\tau,\upsilon} \propto \exp\left[-\frac{\Delta_{\tau,\upsilon}}{\lambda_\theta}\right] \;\text{ where }\; \sum_{\upsilon} \theta_{\tau,\upsilon} = 1 \tag{5}$$

where $\Delta_{\tau,\upsilon} = \min\left[|\tau-\nu|,\, T_{\mathrm{day}} - |\tau-\nu|\right]$, and T_{day} is one day. λ_θ is a hyper-parameter that controls decaying rate in Eq.(5).

2.4.2. Model 1: Principal component analysis (PCA model)

The first model uses global features based on PCA. PCA is a feature extraction technique that linearly transforms the original data to PCs in order of preserving the covariance [10, 14, 31]. In order to obtain the PCs, eigenvalue decomposition is implemented with the regularized empirical covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$.

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\mathbf{y}} &= \frac{1}{N} \sum_{\tau} \sum_{\{t|\mathrm{HoD}(t)=\tau\}} \Delta\mathbf{y}_t \cdot \Delta\mathbf{y}_t^{\mathrm{T}} + \eta^2\, \mathbf{I}_P \\ &= \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathrm{T}} \\ \text{where }\; \Delta\mathbf{y}_t &= \mathbf{y}_t - \hat{\boldsymbol{\mu}}_\tau \end{aligned} \tag{6}$$

N is the total number of timestamps, i.e., $N \simeq N_\upsilon \cdot h$, and the vector $\Delta\mathbf{y}_t$ is the difference between temperature $\mathbf{y}_t$ and its corresponding mean field $\hat{\boldsymbol{\mu}}_\tau$. η is a regularization parameter that has a small value compared to the empirical standard deviation of each location, and $\mathbf{I}_P$ is the identity matrix of dimension $[P \times P]$. The regularization term $\eta^2\, \mathbf{I}_P$ is introduced to avoid overfitting in higher order PCs. Orthogonal matrix $\mathbf{E}$ and (diagonal) eigenvalue matrix $\boldsymbol{\Lambda}$ are outcomes of the eigenvalue analysis on $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$. We determine the dimension of the state variables R by imposing Eq.(7) to reach a small reconstruction error $\hat{\sigma}_{\mathbf{v}}$ that is less than or equal to a target V_{tol}.

$$\hat{\sigma}_{\mathbf{v}}^2 = \frac{1}{P}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}\right] - \frac{1}{P}\sum_{i=1}^{R} \lambda_i \;\le V_{\mathrm{tol}}^2 \tag{7}$$

where λ_i is i-th eigenvalue of the covariance matrix. The dimension R is the number of included PCs. The left-hand side of Eq.(7) estimates the standard deviation $\sigma_{\mathbf{v}}$ of the error term $\mathbf{v}_t$ in Eq.(1). $\mathbf{E}_R$ is the rectangular matrix of the first

R eigenvectors, and $\mathbf{\Lambda}_R$ is the diagonal matrix of the first R eigenvalues. Then, the embedding matrix $\mathbf{\Phi}$ is estimated as follows:

$$\mathbf{\Phi} = \mathbf{E}_R \sqrt{\mathbf{\Lambda}_R} \tag{8}$$

Given the PCA-based embedding matrix $\mathbf{\Phi}$, the expected vector of lower-dimension embeddings $\mathbf{x}_t$ is estimated from vector $\mathbf{y}_t$, by solving a least-square linear regression problem as $\hat{\mathbf{x}}_t = \mathbf{\Phi}^+ \mathbf{y}_t$ where $\mathbf{\Phi}^+$ is Moore-Penrose pseudo inverse matrix of $\mathbf{\Phi}$.

For linear dynamical systems, forward/backward smoothing and sequential optimization, or a sampling-based approach is commonly implemented to calibrate the dynamic model, e.g., expectation-maximization-algorithm or Markov-Chain Monte-Carlo method, as in [28, 3]. However, implementing such conventional algorithms is computationally challenging for our problem given the high dimensionality of the observations and the number of latent variables. Therefore we propose simplified alternative algorithms to calibrate our models. Appendix A investigates the impact of using our methods instead of the conventional ones. The results imply that the the different calibration methods result in similar model performance (see Appendix A for details).

The matrices $\mathbf{F}_\tau$ is estimated by minimizing the following least-squares error function:

$$\mathcal{L}_{\mathrm{LSE}} = \frac{1}{2N} \sum_{\tau} \sum_{\{t|\mathrm{HoD}(t)=\tau\}} \left\| \hat{\mathbf{x}}_t - \mathbf{F}_\tau \hat{\mathbf{x}}_{t-1} \right\|_{L_2}^2 \tag{9}$$

where $\| \cdot \|_{L_2}$ is the Euclidean L^2 norm. To avoid the overfitting, Eq.(9) is modified with a term, which is a variant of ridge penalization, as:

$$\mathcal{L}_{\mathrm{PCA}} = \mathcal{L}_{\mathrm{LSE}} + \frac{1}{2} \alpha_{\mathrm{PCA}} \sum_{\tau} \| \mathbf{F}_\tau - \mathbf{D} \|_{\mathrm{F}}^2. \tag{10}$$

where $\| \cdot \|_{\mathrm{F}}$ is the Frobenius norm, and α_{PCA} is a regularization parameter, which can be selected through n-fold cross validation. $\mathbf{D}$ is a diagonal matrix of which components are the coefficients of uncorrelated uni-variate linear regressions. $\hat{D}_i$, i-th element of the diagonal matrix $\mathbf{D}$, is estimated by multiple uni-variate linear regressions such that

$$\hat{D}_i = \left[\sum_t \hat{x}_t^{(i)} \cdot \hat{x}_{t+1}^{(i)} \right] \left[\sum_t (\hat{x}_t^{(i)})^2 \right]^{-1} \tag{11}$$

where $\hat{x}_t^{(i)}$ is i-th element of $\mathbf{x}_t$. Table 1 summarizes the steps in parameter estimation for the PCA model.

2.4.3. *Model 2: Optimal sensing locations (OS model)*

As discussed in section 2.3, we propose an alternative method for latent variable selection, other than PCA. In the OS model, the latent variables are defined as the residual temperature at the locations from which it would be most informative to measure the temperature field in order to interpolate/extrapolate it entirely. We start appointing the optimal sensing locations by partitioning vector $\Delta\mathbf{y}_t$ of Eq.(6) into two sub-vectors, $\mathbf{x}_t$ and $\mathbf{y}_t^{\mathrm{ip}}$. We define $\mathbf{x}_t$ as the latent state, a vector of residual temperatures at a set of locations $\mathcal{A}$, where it maximizes the information to estimate the rest of field $\mathbf{y}_t^{\mathrm{ip}}$.

$$\Delta\mathbf{y}_t = \mathbf{y}_t - \hat{\boldsymbol{\mu}}_\tau = \begin{bmatrix} \mathbf{x}_t \\ - \\ \mathbf{y}_t^{\mathrm{ip}} \end{bmatrix} \tag{12}$$

The selection of locations $\mathcal{A}$ is identical to the optimal sensor placement problem for a Gaussian random field. Algorithms have been proposed to select the optimal places based on conditional entropy (CE) or mutual information (MI) for arbitrary GPs. This paper adopts the algorithm based on MI, because it tends to place sensors more centrally than CE [15]. The optimal sensor placement $\mathcal{A}^*$ is obtained by solving the following optimization problem.

$$\mathcal{A}^* = \mathrm{argmax}_{\mathcal{A} \subseteq S : |\mathcal{A}| = R} \left[H\left(\mathbf{y}_t^{\mathrm{ip}}\right) - H\left(\mathbf{y}_t^{\mathrm{ip}} | \mathbf{x}_t\right) \right] \tag{13}$$

where $|\mathcal{A}|$ is the number of sensors, which is the dimension of latent variables R. S is a set of locations where sensors can be placed, and one can design the set to shrink the candidate sensing points, e.g. evenly distributed points over the domain. $H(\cdot)$ is the entropy or CE of the random variable(s) within parenthesis, and the simplified forms are provided for multivariate Gaussian random variables as follows:

$$H\left(\mathbf{y}_t^{\text{ip}}\right) = C + \frac{1}{2}\ln(|\mathbf{\Sigma}_{\mathbf{y}^{\text{ip}}}|) \tag{14a}$$

$$H\left(\mathbf{y}_t^{\text{ip}}|\mathbf{x}_t\right) = C + \frac{1}{2}\ln(|\mathbf{\Sigma}_{\mathbf{y}^{\text{ip}}|\mathbf{x}}|) \tag{14b}$$

where C is a constant, and $\mathbf{\Sigma}_{\mathbf{y}^{\text{ip}}}$ and $\mathbf{\Sigma}_{\mathbf{y}^{\text{ip}}|\mathbf{x}}$ are the marginal and conditional covariance matrices of $\mathbf{y}^{\text{ip}}$, given $\mathbf{x}$, respectively. $|\cdot|$ is the determinant of the matrix. The marginal covariance $\mathbf{\Sigma}_{\mathbf{y}^{\text{ip}}}$ in Eq.(14a) is estimated as follows:

$$\hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}} = \frac{1}{N}\sum_t \{\hat{\mathbf{y}}_t^{\text{ip}}\}\cdot\{\hat{\mathbf{y}}_t^{\text{ip}}\}^{\text{T}} + \eta^2\,\mathbf{I}_{\bar{R}} \tag{15}$$

where η is a regularization parameter as in Eq.(6), and $\mathbf{I}_{\bar{R}}$ is the identity matrix of dimension $[(P-R)\times(P-R)]$. Similarly, the conditional covariance $\mathbf{\Sigma}_{\mathbf{y}^{\text{ip}}|\mathbf{x}}$ in Eq.(14b) is estimated using the conditional formula of multivariate Gaussian distribution.

$$\begin{aligned}
&\hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}|\mathbf{x}} = \hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}} - \hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}\mathbf{x}}\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}\mathbf{y}^{\text{ip}}} \\
&\text{where } \hat{\mathbf{\Sigma}}_{\mathbf{x}} = \frac{1}{N}\sum_t \hat{\mathbf{x}}_t\cdot\hat{\mathbf{x}}_t^{\text{T}} + \eta^2\,\mathbf{I}_R \\
&\hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}\mathbf{x}} = \frac{1}{N}\sum_t \hat{\mathbf{y}}_t^{\text{ip}}\cdot\hat{\mathbf{x}}_t^{\text{T}}, \text{ and } \hat{\mathbf{\Sigma}}_{\mathbf{x}\mathbf{y}^{\text{ip}}} = \hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}\mathbf{x}}^{\text{T}}
\end{aligned} \tag{16}$$

where $\mathbf{I}_R$ is the identity matrix of dimension $[R\times R]$. The proposed algorithms in [15] perform a greedy optimization adding new places to sense the temperature until we achieve a satisfactory level of accuracy. One of the merits in using OS model is that the latent variables are directly observable so that no further procedure to estimate latent states is required. The latent states $\mathbf{x}_t$ can be directly estimated from the local measures at the selected sensing locations $\mathcal{A}^*$. The detailed procedure for the optimization in Eq.(13) follows the algorithm in [15]. Reconstruction error $\sigma_{\mathbf{v}}$ determines the dimension of latent variables R, and it is evaluated as the averaged trace of the empirical conditional covariance $\hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}|\mathbf{x}}$.

$$\hat{\sigma}_{\mathbf{v}}^2 = \frac{1}{P-R}\text{Tr}\left[\hat{\mathbf{\Sigma}}_{\mathbf{y}^{\text{ip}}|\mathbf{x}}\right] \le V_{\text{tol}}^2 \tag{17}$$

After having selected the optimal set $\mathcal{A}^*$, the transition matrix is calibrated similarly to how it is done for the PCA model, adopting a least-squared-error (LSE) objective function with a regularization term.

$$\mathcal{L}_{\text{OS}} = \mathcal{L}_{\text{LSE}} + \frac{1}{2}\alpha_{\text{OS}}\sum_\tau \left\|\mathbf{F}_\tau - \mathbf{F}_{\text{HM}}\right\|_{\text{F}}^2, \tag{18}$$

The regularization term is introduced to identify stable transition matrices. $\mathbf{F}_{\text{HM}}$ is a homogeneous and isotropic transition matrix obtained from a pre-training procedure as the solution of the following Markov Gaussian Process.

$$\mathbf{F}_{\text{HM}} = \mathbf{\Sigma}_{\mathbf{x}_t\mathbf{x}_{t-1}}^{\text{HM}}\left[\mathbf{\Sigma}_{\mathbf{x}_t\mathbf{x}_t}^{\text{HM}} + \eta_{\text{HM}}^2\,\mathbf{I}_R\right]^{-1} \tag{19}$$

where $\mathbf{\Sigma}_{\mathbf{x}_t\mathbf{x}_{t-1}}^{\text{HM}}$ and $\mathbf{\Sigma}_{\mathbf{x}_t\mathbf{x}_t}^{\text{HM}}$ are the covariance matrices constructed using the following homogeneous covariance kernel functions, and η_{HM} is a disturbance noise.

$$\mathbf{\Sigma}_{\mathbf{x}_t\mathbf{x}_t}^{\text{HM}}(i,j) = \sigma^2\exp\left[-\frac{\Delta_{i,j}}{\lambda_{\text{HM}}}\right], \tag{20a}$$

$$\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_{t-1}} = \rho\, \mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t} \tag{20b}$$

where $\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t}(i,j)$ is element (i,j) of matrix $\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t}$. σ and ρ are scalar parameters that represent the standard deviation and correlation coefficient, respectively. λ_{HM} is a parameter of the exponentially decaying function in Eq.(20a), and $\Delta_{i,j}$ is the geometric distance between sites i and j. σ, ρ, and λ_{HM} are calibrated through a pre-training procedure by maximizing log-likelihood function as:

$$\left[\hat{\sigma},\, \hat{\rho},\, \hat{\lambda}_{\mathrm{HM}}\right] = \mathrm{argmax}\; LL\left(\{\mathbf{x}_t, \mathbf{x}_{t-1}\}; \mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t}, \mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_{t-1}}\right) \tag{21}$$

where $LL(\cdot)$ is the log-likelihood function of Gaussian distribution for the given pair of data $\{\mathbf{x}_t, \mathbf{x}_{t-1}\}$ and homogeneous covariance, $\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t}$ and $\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_{t-1}}$. After calibrating the pre-training parameters, the regularization coefficient α_{OS} is then selected through n-fold cross validation.

In the OS model, we estimate the embedding matrix through linear regression. Similar to Eq.(18), linear regression is implemented with the homogeneous regularization scheme. The embedding matrix is the optimal solution of the following loss function.

$$\mathcal{L}^{\Phi}_{\mathrm{OS}} = \mathcal{L}^{\Phi}_{\mathrm{LSE}} + \frac{1}{2}\alpha_{\mathrm{OS}} \|\mathbf{\Phi} - \mathbf{\Phi}_{\mathrm{HM}}\|^2_{\mathrm{F}} \tag{22}$$

where

$$\mathcal{L}^{\Phi}_{\mathrm{LSE}} = \frac{1}{2N}\sum_t \left\|\Delta\mathbf{y}_t - \mathbf{\Phi}\,\mathbf{x}_t\right\|^2_{L_2}. \tag{23}$$

The hyper-parameter α_{OS} in Eq.(22) is set to be equal to that of Eq.(18). The homogeneous embedding matrix $\mathbf{\Phi}_{\mathrm{HM}}$ is defined similarly to the homogeneous transition matrix $\mathbf{F}_{\mathrm{HM}}$ sharing the the parameter λ_{HM} and σ.

$$\mathbf{\Phi}_{\mathrm{HM}} = \mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{y}_t\mathbf{x}_t}\left[\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t} + \eta^2_{\mathrm{HM}}\mathbf{I}_R\right]^{-1} \tag{24}$$

where $\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{y}_t\mathbf{x}_t}$ is a rectangular matrix of which elements is similarly defined as $\mathbf{\Sigma}^{\mathrm{HM}}_{\mathbf{x}_t\mathbf{x}_t}$ in Eq.(20). Table 2 summarizes the details in parameter estimation for the OS model.

2.4.4. Noise covariance

The process noise is calculated with the estimated parameters in Section 2.4.2 (PCA model) or Section 2.4.3 (OS model).

$$\hat{\mathbf{w}}_t = \hat{\mathbf{x}}_{t+1} - \hat{\mathbf{F}}_\tau\, \hat{\mathbf{x}}_t \tag{25}$$

The moving average scheme in Eq.(4) is also applied to estimate time-dependent covariance matrix $\mathbf{\Sigma}^{\tau}_{\mathbf{w}}$ sharing the parameter $\theta_{\tau,\upsilon}$ and λ_θ of Eq.(4) and Eq.(5).

$$\hat{\mathbf{\Sigma}}^{\tau}_{\mathbf{w}} = \sum_{\upsilon}\theta_{\tau,\upsilon}\,\bar{\mathbf{\Sigma}}^{\upsilon}_{\mathbf{w}} \;\text{ where }\; \bar{\mathbf{\Sigma}}^{\upsilon}_{\mathbf{w}} = \frac{1}{N_\upsilon}\sum_{\{t|\mathrm{HoD}(t)=\upsilon\}}\hat{\mathbf{w}}_t\hat{\mathbf{w}}^{\mathrm{T}}_t \tag{26}$$

In Eq.(26), $\hat{\mathbf{\Sigma}}^{\tau}_{\mathbf{w}}$ is the estimated covariance matrix of $\mathbf{w}_t$ at τ while $\bar{\mathbf{\Sigma}}^{\tau}_{\mathbf{w}}$ is the empirical covariance matrix.

2.4.5. Stationary distribution

If the calibrated system is stationary, the process converges to its stationary distributions as the process progresses. Based on the diurnal cycle, the process becomes time-invariant with daily equivalent notations, as described below:

$$\mathbf{x}_t = \mathbf{F}_{\mathrm{D}}\,\mathbf{x}_{t-h} + \mathbf{w}_t{}^{\mathrm{D}}, \;\text{ where }\; \mathbf{F}_{\mathrm{D}} = \prod_\tau \mathbf{F}_\tau \tag{27}$$

where $\mathbf{F}_{\mathrm{D}}$ is the one-day equivalent transition matrix, and $\mathbf{w}^{\mathrm{D}}_t$ is the equivalent process noise. The cumulative noise covariance $\mathbf{\Sigma}^{\mathrm{D},\tau}_{\mathbf{w}}$ of $\mathbf{w}^{\mathrm{D}}_t$ is a function of τ to indicate the time of the day of the stationary distribution. The state covariance $\mathbf{\Sigma}^{\tau}_{\mathbf{x}}$ converges to its stationary point with the following Lyapunov equation:

$$\mathbf{\Sigma}^{\tau}_{\mathbf{x}} = \mathbf{F}_{\mathrm{D}}\mathbf{\Sigma}^{\tau}_{\mathbf{x}}\mathbf{F}^{\mathrm{T}}_{\mathrm{D}} + \mathbf{\Sigma}^{\mathrm{D},\tau}_{\mathbf{w}}. \tag{28}$$

We set this stationary distribution of the process as an estimator for the initial distribution of the system when implementing probabilistic forecast using Kalman Filter/Smoother scheme in Section 2.5.

Table 1
Procedure to estimate parameters (PCA model)

Step	Procedure	Parameter
1	Estimate the mean field $\boldsymbol{\mu}_\tau$ as in Eq.(4)	$\boldsymbol{\mu}_\tau$
2	Estimate latent variable $\mathbf{x}_t$ using Eq.(6). Eq.(7) and Eq.(8) determine the error standard deviation $\sigma_\mathbf{v}$ and the embedding matrix $\boldsymbol{\Phi}$	$\mathbf{x}_t$, $\boldsymbol{\Phi}$, $\sigma_\mathbf{v}$
3	Implement uni-variate linear regressions, as in Eq.(11), to identify the diagonal matrix $\mathbf{D}$ in Eq.(10)	$\mathbf{D}$
3	Determine the regularization hyper-parameter α_{PCA} in Eq.(10) with 10-fold cross validation	α_{PCA}
4	Calibrate the transition matrix $\mathbf{F}_\tau$ minimizing Eq.(10)	$\mathbf{F}_\tau$
5	Estimate the noise covariance $\boldsymbol{\Sigma}^\tau_\mathbf{w}$ as Eq.(26)	$\boldsymbol{\Sigma}^\tau_\mathbf{w}$
6	Evaluate the stationary covariance $\boldsymbol{\Sigma}^\tau_\mathbf{x}$ through Eq.(28)	$\boldsymbol{\Sigma}^\tau_\mathbf{x}$

Table 2
Procedure to estimate parameters (OS model)

Step	Procedure	Parameter
1	Estimate the mean field $\boldsymbol{\mu}_\tau$ as in Eq.(4)	$\boldsymbol{\mu}_\tau$
2	Estimate latent variable $\mathbf{x}_t$ with optimization Eq.(13). Eq.(17) determines the error standard deviation $\sigma_\mathbf{v}$.	$\mathbf{x}_t$, $\sigma_\mathbf{v}$
3	Using the covariance kernel as in Eq.(20), find the parameters that maximize the log-likelihood in Eq.(21)	σ, ρ, λ_{HM}
4	Determine the regularization hyper-parameter α_{OS} in Eq.(18) with 10-fold cross validation	α_{OS}
5	Calibrate the transition matrix $\mathbf{F}_\tau$ minimizing Eq.(18)	$\mathbf{F}_\tau$
6	Calibrate the embedding matrix $\boldsymbol{\Phi}$ minimizing Eq.(22) with α_{OS} of Step 4	$\boldsymbol{\Phi}$
7	Estimate the noise covariance $\boldsymbol{\Sigma}^\tau_\mathbf{w}$ as Eq.(26)	$\boldsymbol{\Sigma}^\tau_\mathbf{w}$
8	Evaluate the stationary covariance $\boldsymbol{\Sigma}^\tau_\mathbf{x}$ through Eq.(28)	$\boldsymbol{\Sigma}^\tau_\mathbf{x}$

2.4.6. Stability, stationarity, and identifiability

A discrete linear process $\mathbf{x}_{t+1} = \mathbf{A}\,\mathbf{x}_t$ is called asymptotically stable if all the eigenvalues of the transition matrix $\mathbf{A}$ have modulus less than one. Similarly, in our models, we have analyzed eigenvalues of the one-day equivalent transition matrix $\mathbf{F}_D$ to confirm that the calibrated process is asymptotically stable. According to our analysis, the maximum values of the modulus ranges from 0.66 to 0.78 for one-day equivalent processes. On the other hand, one may impose the stability condition by parameterizing the transition matrix during the model calibration stage, as in [5].

Strictly speaking, the proposed model is not stationary at each time of the day, but it is stationary over days, as the model structure defines a periodic process (cyclostationary process). For the detailed discussion, readers may refer to [11].

Lastly, the proposed models are identifiable, and the calibrated model parameters are unique. In general, a linear dynamical system, or linear Gaussian state-space model, cannot guarantee identifiability because of the redundancies in the solution space. However, by imposing the PCA/OS model to define latent variables, the solution space is constrained such that either of the proposed models is identifiable. We refer the reader to Chapter 24 of [3] for details on this issue.

2.5. Probabilistic forecast

The temperature field $\mathbf{y}_t$ at t, its prior distribution follows the Gaussian distribution as:

$$\mathbf{y}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{Y}_t},\ \boldsymbol{\Sigma}_{\mathbf{Y}_t}\right) \tag{29}$$

where $\boldsymbol{\mu}_{\mathbf{Y}_t}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}_t}$ are the prior mean and covariance of spatial-temporal temperature field of the target domain. Given the linear combination as Eq.(3), the measurement vector $\mathbf{z}_t$ also follows Gaussian distribution, and its mean $\boldsymbol{\mu}_{\mathbf{z}_t}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{z}_t}$ are:

$$\mathbf{z}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{Z}_t},\ \boldsymbol{\Sigma}_{\mathbf{Z}_t}\right),\ \text{where}\ \ \boldsymbol{\mu}_{\mathbf{Z}_t} = \mathbf{H}_t\,\boldsymbol{\mu}_{\mathbf{Y}_t},\ \ \boldsymbol{\Sigma}_{\mathbf{Z}_t} = \mathbf{H}_t\boldsymbol{\Sigma}_{\mathbf{Y}_t}\mathbf{H}_t^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}. \tag{30}$$

Denoting the observation $\mathbf{z}_{1:T}$ to be a collection of $\mathbf{z}_t$ for $t = 1,\ 2,\ 3,\ \ldots T$, forecasting near-surface temperature fields is to infer the field vector $\mathbf{y}_t$ with the given observations $\mathbf{z}_{1:T}$.

$$\mathbf{y}_t|\mathbf{z}_{1:T} \equiv \mathbf{y}_{t|1:T} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}_t|1:T},\ \boldsymbol{\Sigma}_{\mathbf{y}_t|1:T}\right) \tag{31}$$

where $\boldsymbol{\mu}_{\mathbf{y}_t|1:T}$ and $\boldsymbol{\Sigma}_{\mathbf{y}_t|1:T}$ are posterior mean and covariance, respectively. For state-space models, the inference procedure can be implemented by a recursive Gaussian linear smoother, also known as a Kalman Filter/Smoother, following the details described in [3].

3. Model calibration

This section describes how we have collected data and calibrated the proposed models.

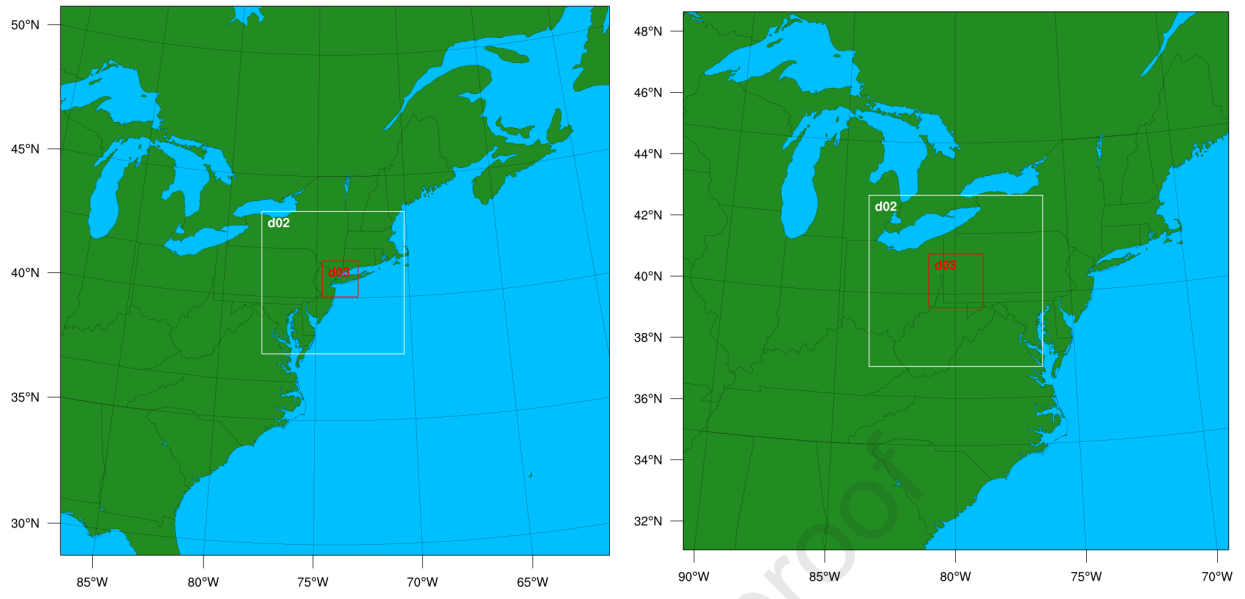
3.1. Weather data

High-resolution numerical simulations have been increasingly used to study the dynamics in urban-scale weather effect and to mitigate related risk [12, 18, 17]. The temporal resolution of interest is also related to the urgency of decision framework for policy makers, such as air/water quality control, construction project design, and energy/groundwater use [9, 17]. In this paper, we have designed 1km grid spacing as the spatial resolution adopting the one-way nesting domains, as in [17, 19]. The spatial resolution would be computationally affordable for data collection, without expecting significant systematic error in local-scale physics of the numerical simulator, but that resolution would be still challenging for real-time computation. We also have set 30-minutes as the reporting intervals, considering that this temporal resolution is tolerable for the potential urgency of decision-making problems.

To calibrate the model, historical weather data over the target domains are reanalyzed, using WRF-PUCM, by downscaling 6-hourly 12km×12km data of North American Meso-scale Forecast System (NAM) 12km Analysis (NAM-ANL; [2]) to 1km×1km grid data with 30-minutes intervals. Using the dynamically downscaled data, we have calibrated the proposed models for both domains around NYC (160km×160km) and PGH (200km×200km) as represented in Figure 2. The weather simulation is conducted with a high-performance computer, Cheyenne, of University Corporation for Atmospheric Research, using 36 cores (72 logical processors) of 2.3-GHz Intel Xeon E5-2697V4 (Broadwell) processors. The simulation data of this paper is published in [7].

3.2. Training

As a result of the WRF-PUCM, a sequence of temperature fields is generated from Jun-01 (UDT+00:00) to Aug-31 (UDT+00:00) with 30-minutes intervals for three summers from 2016 to 2018 in both NYC and PGH domains (1km×1km resolution). A total of 13,251 stamps of temperature fields are used to calibrate the proposed models as in Section 2. Table 3 summarizes the dataset that we generated and used for the calibration. Following the calibration procedures as summarized in Table 1 and Table 2, we have estimated the parameters of the proposed models. Figure 3a scatters the collected data and plots the calibrated mean temperature with 95% confidence bounds in a cycle of a day, at Brooklyn, NYC (40.668°N, 73.994°W). Figure 3b represents the histogram of near-surface temperature and its fitted probability distribution of the OS model at 18:00 for the same location. Table 4 and Table 5 list the hyper-parameters and their selected values in the model calibration. The tolerance value V_{tol} is differently selected in Table 4 because the PCA model shows more rapidly decreasing reconstruction error, even with fewer latent variables, than the OS model.



(a) New York City **(b)** Pittsburgh

Figure 2: Geographical representation of the modeling domains (d03, red box)

Table 3
Summary of training data

Setting	New York City	Pittsburgh
Latitude	[40.040, 41.513]° N	[39.546, 41.325]°N
Longitude	[74.701, 72.741]°W	[81.172, 78.777]°W
Spatial grid	159×159	201×198
Grid spacing	1km	1km
Calibration years	2016/2017/2018	2016/2017/2018
Start date	Jun-01 00:00 UTC	Jun-01 00:00 UTC
End date	Sep-1 00:00 UTC	Sep-1 00:00 UTC
Time interval	30 min	30 min

Table 4
Selected hyper-parameters I

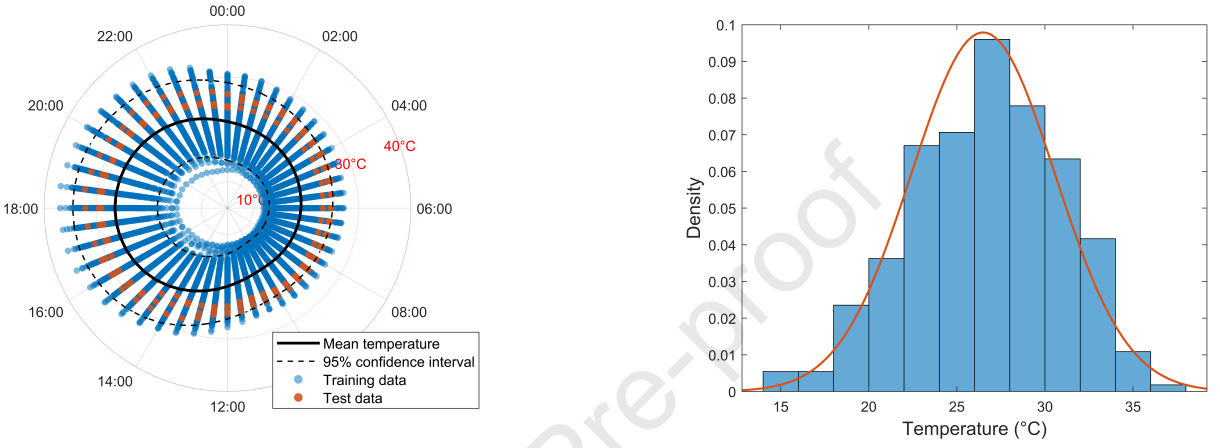
Hyper-parameter		PCA model	OS model
λ_θ	[hour]	0.50	0.50
η	[°C]	0.32	0.32
V_{tol}	[°C]	0.32	0.44
η_{HM}	[°C]	-	0.32

4. Model validation

In this section, we conduct a numerical experiment for the calibrated models to forecast short-term near-surface temperature at relatively low computational cost. In Section 4.1, we provide realistic model case studies and show the results. Then, the prediction performance is assessed by comparing the accuracy and the evaluation time in Section 4.2.

Table 5
Selected hyper-parameters II

Hyper-parameter		Model	New York City	Pittsburgh
α_{PCA}	[-]	PCA model	0.01	3.03
σ	[°C]	OS model	3.05	3.43
ρ	[-]	OS model	0.99	0.98
λ_{HM}	[km]	OS model	15.35	15.10
α_{OS}	[-]	OS model	0.20	0.35



(a) Calibrated daily temperature

(b) Fitted distribution (18:00 UTC)

Figure 3: Fitted distributions of the calibrated OS model (Brooklyn, New York City)

4.1. Short-term forecast

The meaning of "short-term" depends on the purpose of the forecast or analysis. Among the officially operated weather forecasting systems, the North American Meso-scale Forecast System Non-hydrostatic Mesoscale Model (NAM-NMM) is one of the most detailed models, which provides 3-hourly 12km×12km grid weather forecasts for the next 84-hours, up to the knowledge of the authors. Referring to the NAM and considering the computational trade-off in spatial-temporal resolution, we assume that a faster forecast predicts the next a few hours to a day at 1km×1km spatial resolution, with tens of minutes to an hourly time intervals.

For the numerical test, the calibrated models in Section 3 predict the downscaled near-surface temperature of WRF-PUCM from Aug-01-2019 to Aug-03-2019, which is at least one year apart from the training period. We set input data to the calibrated model, assuming them as noisy signals with zero-mean Gaussian white noise of which standard deviation is 0.1°C, i.e., $\boldsymbol{\epsilon}_t$ in Eq.(3); the noise level is set for the proposed model to provide a fair comparison to WRF-PUCM, not being disturbed by input noise. We consider the following use cases to implement the fast forecast:

Case 1: we assume an exploitation of local temperature measurements on the virtual m weather stations $\{y_t^{(1)}, y_t^{(2)}, ..., y_t^{(m)}\}$. In this case, the simulation outcomes of WRF-PUCM, as collected field measurements, are accessible to the models up to the current time step. The weather stations are assumed to be evenly located with 6km spacing.

Case 2: In addition to Case 1, average temperature over the domain (global average) is provided, from WRF-PUCM data, with 6-hourly intervals for the test period. This case assumes that the availability of the broader scale temperature forecast from an external source, which has sparse temporal intervals and coarse spatial resolution, similarly to the test case of [19].

Case 3: This case provides multiple average temperatures of 12km ×12km grids, as shown in Figure 4d, from WRF-PUCM data. This case describes a realistic scenario, which exploits the most information among the case studies. One can consider the multiple zonal averages as a prediction by NWP simulation at coarse and sparse resolution. The grid spacing and time intervals are set to be similar to those of NAM.

Table 6
Summary of the model case studies

	Field measurement		External forecast	
	Grid spacing	Time interval	Zonal scale	Time interval
Case 1	6 km	30 min	-	-
Case 2	6 km	30 min	Domain size	6 hourly
Case 3	6 km	30 min	12 km × 12 km	6 hourly
Baseline 1	6 km (< 10 km)	up to 24 hours	-	-
Baseline 2	-	-	12 km × 12 km	6 hourly
Baseline 3	6 km (< 10km)	up to 24 hours	12 km × 12 km	6 hourly

To assess the performance of the proposed models, we set baselines as the various uses of the GP model, which was developed in [19]. However, because of the computational complexity of the fully-connected GP, we set constraints, by limiting the number of accessible data points, to attain the temperature prediction with reasonable computing resources.

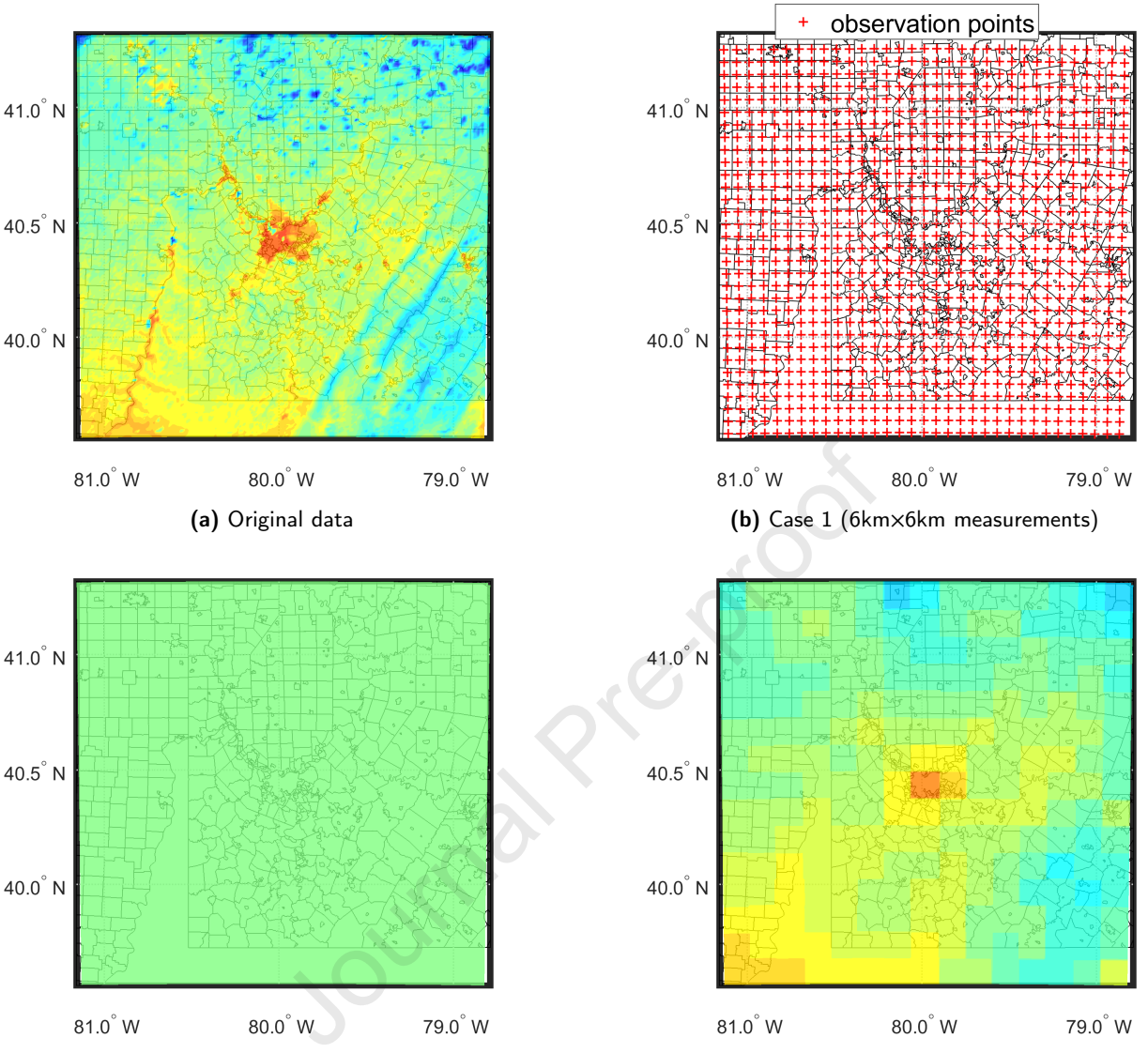
Baseline 1: this baseline forecasts near-surface temperature only using the past measurements as in Case 1, but with the limited data. We force the model to forecast near-surface temperature of individual locations, only exploiting the measurements within a radius of 10 km sub-domain at selected time points, up to the last 24-hours, {Current time − [00:00, 00:30, 01:00, 03:00, 06:00, 09:00, 12:00, 18:00, 24:00]}. This case is the closest configuration to Case 1 under the limited budget of computation.
Baseline 2: For the test period, we input the 12km×12km zonal averages only. This case is the fastest implementation that feeds overall spatio-temporal information to the GP model.
Baseline 3: In addition to Baseline 1, Baseline 3 uses the 6-hourly 12km×12km average temperature, but the individual location is allowed to exploit the zonal average where it belongs. This baseline is considered as the equivalent use case to Case 3, but with the constrained computing resource.

Table 6 summarizes the information used in the case studies and the baselines. Other than spatio-temporal models, we also consider much simpler ways to forecast near-surface temperature. We evaluate the following simple quantities to investigate whether the probabilistic models enable improved temperature forecast.

Simple forecast 1 $T_{t-24\mathbf{h}}$: The temperature is predicted by the temperature of the previous day at the closest virtual weather station of Case 1.
Simple forecast 2 $T_{\mathbf{avg}}$: The temperature is predicted by the mean temperature of the location at the same hour of the day τ.

Figure 5a represents the reanalyzed near-surface temperature by WRF-PUCM on Aug-02-2019 19:00 UTC (15:00 EDT) for the NYC domain, which is assumed as ground truth for this numerical experiment. Figure 5b illustrates the 24-hours ahead prediction by the baseline 3 for the same time point. Figure 5c and Figure 5d are the 24-hours ahead prediction by the proposed state-space models. Similarly, Figure 6a represents the reanalyzed near-surface temperature by WRF-PUCM on Aug-03-2019 19:30 UTC (15:30 EDT) for the PGH domain, and Figure 6b, Figure 6c, and Figure 6d are the 24-hours ahead prediction by the baseline 3 and the use of Case 3 with the proposed models. The selected time points are when each domain shows the hottest average temperature. In Figure 6a, the spots with low temperature, which do not appear in the proposed models, are predicted cloudy areas by the NWP model. Although cloud positions have been hard-to-predict quantities for NWP models, we do not investigate the exact cloud position because it is out of the scope of the paper. The comparison would be still valid as the model performance targets WRF-PUCM. Because no cloud effect has been considered in the model, the result can be comprehended as a trade-off between faithful physical description and fast forecast. However, one may see that the proposed models still perform well for the overall regions, showing high accuracy, in Section 4.2. Figure 7 shows the 3-hours ahead predicted temperature by the OS model during the test periods at Brooklyn, NYC, and Downtown, PGH.

4.2. Model assessment



(a) Original data

(b) Case 1 (6km×6km measurements)

(c) Case 2 (global average)

(d) Case 3 (12km×12km grid average)

Figure 4: Illustration of input data for case studies

To assess the performance of the models, we conduct two types of validation: (1) point-wise and (2) probabilistic validation. In both cases, we compare pairs of observations and mean-predictions. The point-wise validation assesses model performance by comparing this pair using traditional error metrics, e.g., root-mean-squared error. On the other hand, the probabilistic validation assesses the performance of probabilistic forecasts by comparing the probability distributions of the pair. For example, we evaluate the probability of observations to be drawn from the posterior temperature distribution as a score of the probabilistic model assessment.

4.2.1. Point-wise validation

We use the following error metrics: (1) the root-mean-squared error (RSME), (2) the unnormalized bias (b_{bias}), (3) the coefficient of variation in the mean absolute error (CvMAE), and (4) the Pearson linear correlation coefficient (r).

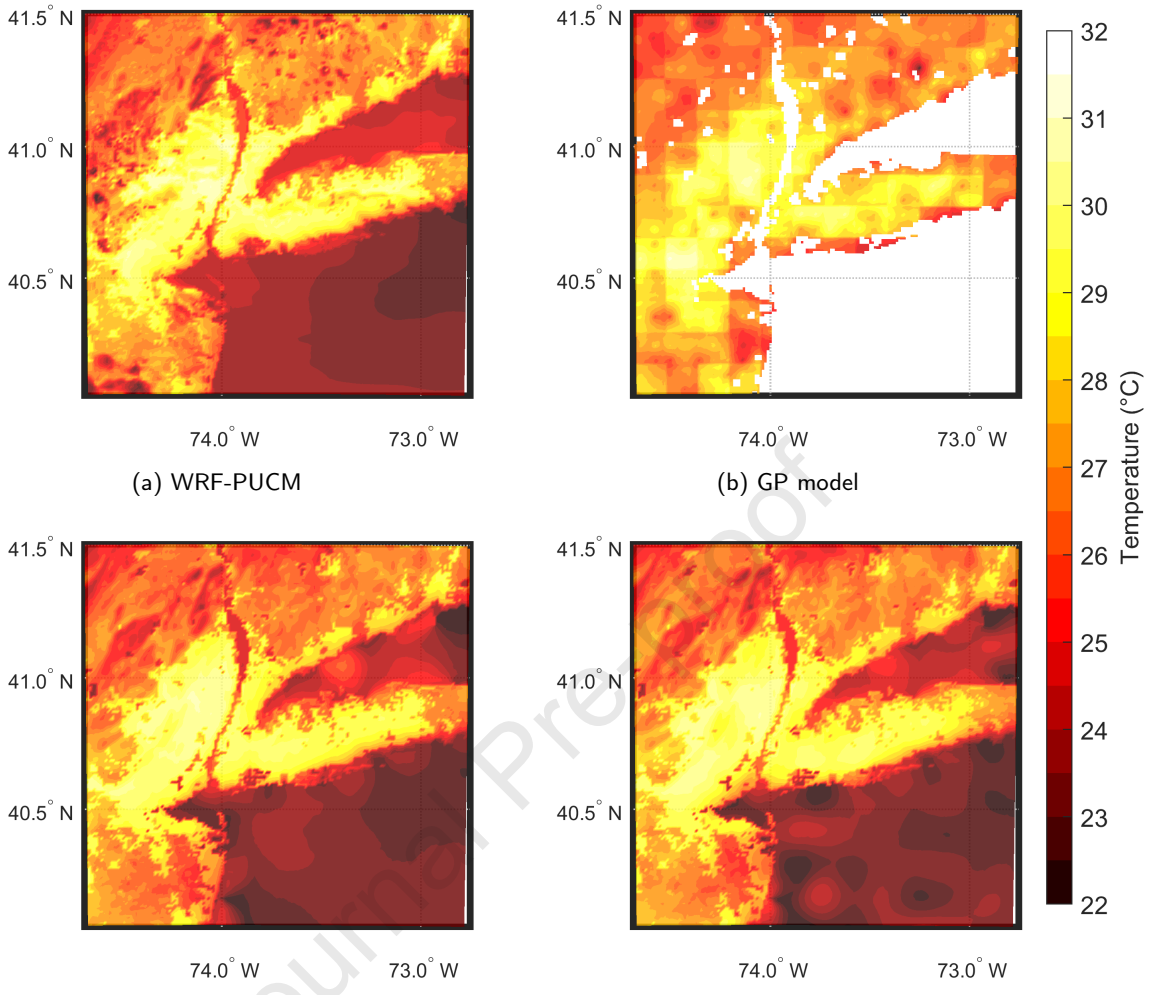


Figure 5: (a) near-surface temperature at Aug-02-2019 19:00 UTC (15:00 EDT), (b) 24-hours ahead prediction by the GP model (Baseline 3), (c) the PCA model (Case 3), and (d) the OS model (Case 3)

The last three metrics are adopted from [21]. Denoting the prediction outcomes of the WRF-PUCM and the proposed models as y_i and $\hat{y}_i$, respectively, the RSME is defined as the following:

$$\text{RSME} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{32}$$

where n is the number of the target and prediction pairs. The forecast bias is assessed by b_{bias} and defined as the average difference between the predictions and the actual values.

$$b_{\text{bias}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i). \tag{33}$$

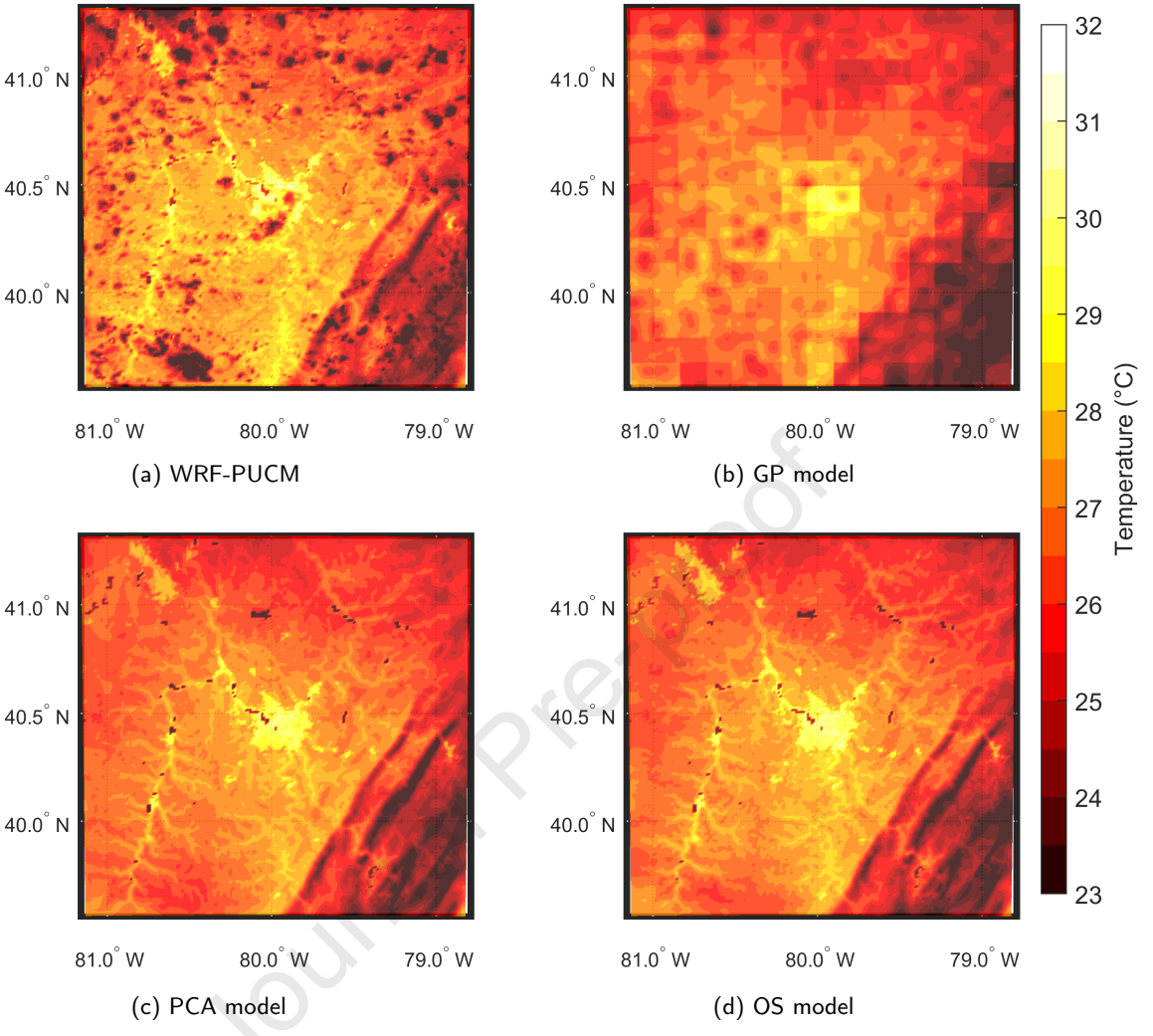


Figure 6: (a) near-surface temperature at Aug-03-2019 19:30 UTC (15:30 EDT), (b) 24-hours ahead prediction by the GP model (Baseline 3), (c) the PCA model (Case 3), and (d) the OS model (Case 3)

CvMAE computes the deviation of the predicted values as the normalized absolute errors against the bias.

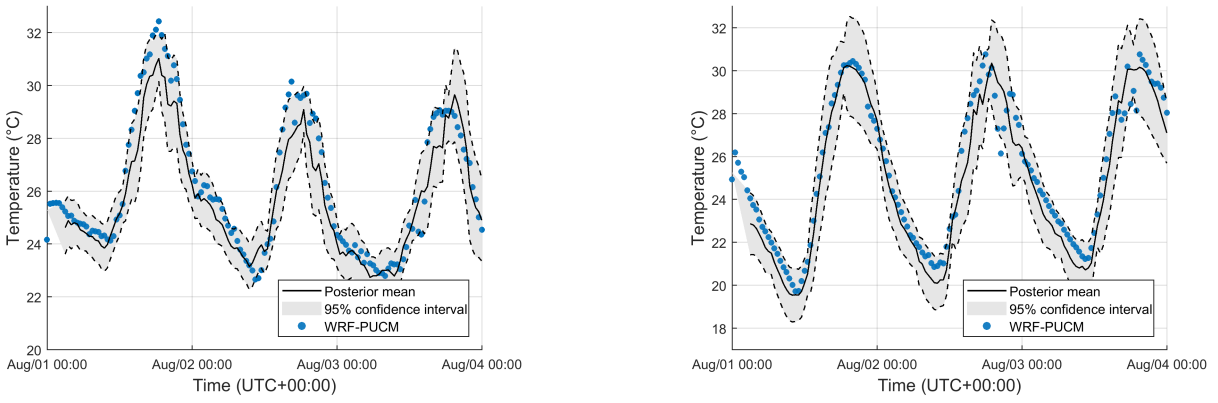
$$\text{CvMAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - b_{\text{bias}} - y_i|}{\sum_{i=1}^{n} y_i} \tag{34}$$

Lastly, the Pearson linear correlation coefficient (r) assesses the linearity of the predictions to the actual values.

$$r = \frac{\sum_{i=1}^{n} \Delta\hat{y}_i \cdot \Delta y_i}{\sqrt{\sum_{i=1}^{n} \Delta\hat{y}_i{}^2} \sqrt{\sum_{i=1}^{n} \Delta y_i{}^2}} \tag{35}$$

where

$$\Delta y_i = y_i - \frac{1}{n} \sum_{j=1}^{n} y_j. \tag{36}$$



(a) Brooklyn, New York City **(b)** Downtown, Pittsburgh

Figure 7: 3-hours ahead prediction by OS model, Aug-01-2019 to Aug-04-2019 (UTC+00:00)

4.2.2. Probabilistic validation

We consider the continuous-ranked-probability score (CRPS) and the cross-entropy (CE) as error metrics for probabilistic validation. CRPS is defined as follows:

$$\mathrm{CRPS}(F_{\hat{y}_i}, y_i) = \int_{-\infty}^{\infty} \left[F_{\hat{y}_i}(\hat{y}_i) - \mathbb{I}(\hat{y}_i - y_i) \right]^2 d\hat{y}_i \tag{37}$$

where $F_{\hat{y}_i}$ is the cumulative density function of the prediction $\hat{y}_i$, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if its input $(\hat{y}_i - y_i)$ is positive or zero, or 0 otherwise. CPRS measures the divergence between the cumulative density functions of probabilistic forecast and its corresponding observation.
Along with CRPS, we also consider CE to measure the divergence between probability densities of two continuous random variables, as follows:

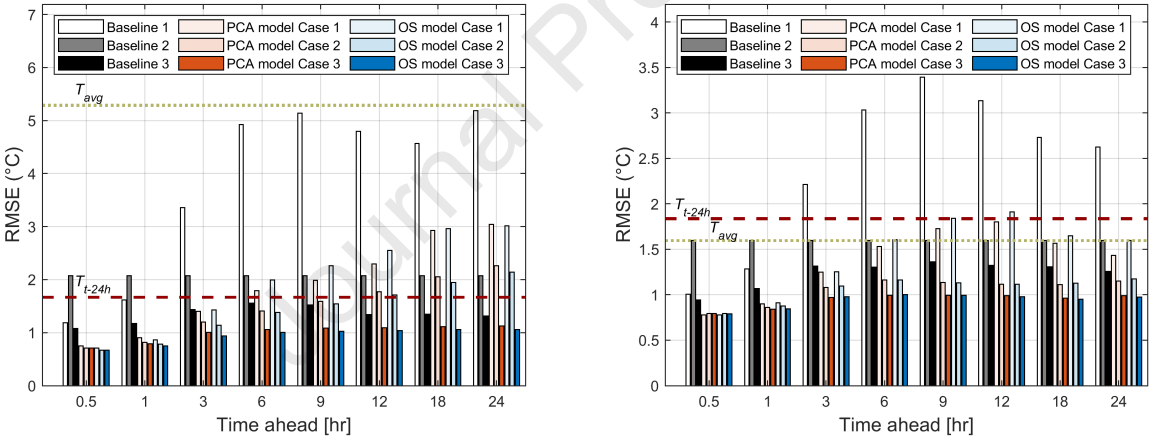
$$\begin{aligned} \mathrm{CE}(f_{y_i}, \hat{f}_{y_i}) &= -\int_{-\infty}^{\infty} f_{y_i}(x) \log f_{\hat{y}_i}(x) dx \\ &= -\int_{-\infty}^{\infty} \delta_{y_i}(x - y_i) \log f_{\hat{y}_i}(x) dx \\ &= -\log \hat{f}_{y_i}(y_i) \end{aligned} \tag{38}$$

where f_{y_i} and $\hat{f}_{y_i}$ are probability density functions of the observation and the corresponding prediction, respectively. In Eq.(38), we have replaced f_{y_i} with the shifted Dirac-delta function $\delta(x - y_i)$ since the true distribution of the observation is unknown. Thus, CE is equivalent to the negative log-likelihood of the observation to be drawn from the prediction distribution.

4.2.3. Results

For the different settings on the prediction lead time (00:30, 01:00, 03:00, 06:00, 09:00, 12:00, 18:00, 24:00), the forecast results are assessed under the case studies. Figure 8 illustrates RSME with the prediction lead time and also plots the errors of the simple forecasts. In most case studies, the proposed models showed reduced forecast errors than the baselines and the simple cases. Recalled that Case 1 and 2 use much less information, Case 3 shows a considerable capacity to predict the near-surface temperature, achieving the accuracy around, or less than, 1°C RSME in both domains. According to the results, Simple forecast 1, $T_{t-24\mathrm{h}}$, outperforms several use scenarios of the PSTMs. However, this simple forecast never reaches the probabilistic models' best performance that leverages the spatio-temporal patterns to forecast near-surface temperature using the various input information; its prediction error, 1.67-1.87°C, is far above that of the proposed models (Case 3), 0.97-1.13°C. The result reveals that Simple forecast 2, T_{avg}, is not a reliable predictor because it performs well in the PGH domain while it is the worst predictor in the NYC domain; this measure

does not work well, especially when the current weather is quite dissimilar to the average.
GP's inherent weakness in extrapolation would explain the unsatisfactory performance of Baseline 1. Baselines 1 and 3 show trends that they reduce the prediction error as the prediction lead time approaches 24-hours, except for 24-hours ahead prediction of Baseline 1 in NYC. It would be because the GP model uses a periodic kernel function with a daily cycle. In the Figure 8b, as the prediction lead time increases, the estimated error of Case 1 approaches, passes over, and comes back to Simple forecast 2 T_{avg}, both in PCA and OS model. The behavior represents the characteristics of the proposed models. Given the calibrated one-day equivalent transition $\mathbf{F}_{\text{D}}$, of which all the absolute eigenvalues are less than 1, a random state of the system will eventually converge to the stationary distribution related to Eq.(28) with the significant temporal progress.
In Case 2, the prediction is more precise in the PGH domain than the NYC domain. The result implies that the global average, the input of the case study, can estimate the latent states well in the PGH region, but the NYC region requires more detailed information to identify the system state due to the higher uncertainty. This remark is consistent with the higher temperature variance in NYC. For example, in NYC, Brooklyn shows a 4.07°C standard deviation at 18:00 UTC (14:00 EDT), having its highest mean temperature 26.49°C, while Downtown, PGH, has the highest mean value 28.45°C at 19:30 UTC (15:30 EDT) but a smaller standard deviation of 3.09°C. Possible reasons for the higher variance can be the land use, i.e., high urbanization, and the influence of the Atlantic Ocean. The NYC domain contains 24.38% of grid cells classified as an urban category, of which the most dominant land use is urban, while 37.17% of the region is water surfaces (mostly oceans). In contrast, the PGH domain includes only 3.23% and 0.51% of urban area and water surfaces (mostly rivers), respectively. However, a more sophisticated study should be required to identify the sources of temperature variability.



(a) New York City **(b)** Pittsburgh

Figure 8: Root-mean-squared error (RSME) by prediction lead time. In this figure, prediction errors of individual locations are computed by comparing pairs of simulated data and probabilistic forecasts at different prediction lead times, then RSME is evaluated with the collected prediction errors over the entire domain. The dashed lines represent the results of the simple forecasts $T_{t-24\text{h}}$ and T_{avg}

Table 7 and Table 8 provides the 24-hours ahead forecast RSME for the different land uses. GP model is set to forecast near-surface temperature, excluding water surfaces (e.g., ocean) from the NYC domain. Due to its homogeneous covariance kernels, the GP model may predict land temperature erroneously when it considers the wetland temperature. For example, because of the homogeneous covariance kernels, two local temperatures at land and ocean would have the same correlation coefficient with the temperature at a third location, in case that the geometric distances from two locations to the third are equal, which yields erroneous forecasts. Non-urban areas denote grid cells where the most dominant use type is neither urban nor wetland.
For both NYC and PGH, the proposed model shows better performance in most cases; even though Baseline 3 is the best predictor for urban areas around NYC, the proposed models also show almost equal performance in Case 3.

Table 7
Root-mean-squared error by land use, New York City

RSME (°C)	GP model			PCA model			OS model		
	Baseline 1	Baseline 2	Baseline 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
Urban	4.00	2.27	0.96	2.97	1.35	1.07	3.21	1.33	0.98
Wetland	-	-	-	2.62	2.76	0.74	2.36	2.59	0.65
Non-urban	5.83	1.98	1.49	3.44	2.19	1.43	3.42	2.09	1.38
Overall	5.19	2.10	1.31	3.04	2.26	1.13	3.01	2.14	1.06

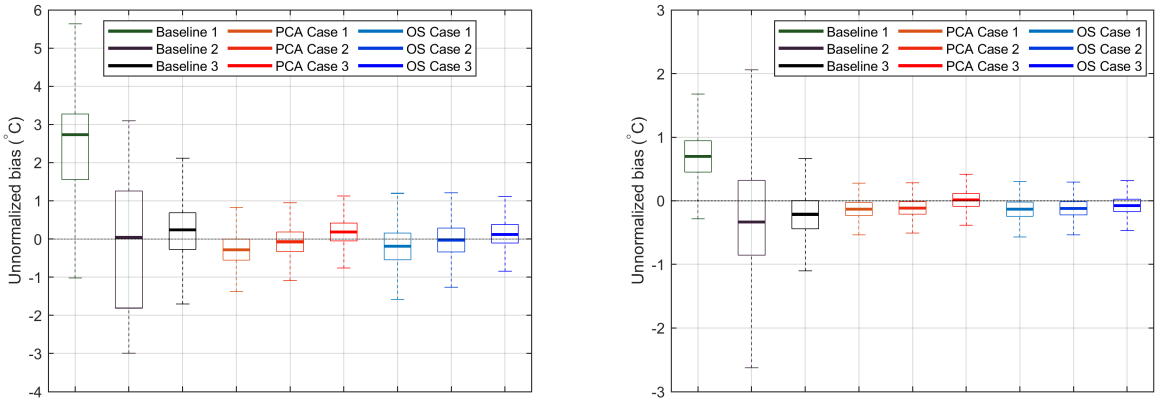
Table 8
Root-mean-squared error by land use, Pittsburgh

RSME (°C)	GP model			PCA model			OS model		
	Baseline 1	Baseline 2	Baseline 3	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
Urban	1.31	3.30	1.71	1.32	1.00	0.84	1.44	1.01	0.84
Wetland	2.50	2.10	1.93	2.20	1.79	1.30	2.21	1.70	1.39
Non-urban	2.66	1.49	1.23	1.43	1.15	0.99	1.60	1.17	0.97
Overall	2.62	1.59	1.26	1.43	1.15	0.99	1.60	1.17	0.97

Overall prediction performance is better in PGH than in NYC, and it might imply more predictable weather patterns in the PGH domain, having less variability as discussed in the previous paragraph. In comparing urban and non-urban areas, urban areas tend to hold higher predictability in both domains except Baseline 2. Possible rationale includes the central location of urban areas in the test domains. Because the spatial correlation tends to decay with the geometric distance, both PCA and OS models may have learned centered temperature patterns over the domain. The wetland temperature in NYC is the most predictable quantity for Case 3, while it is the most challenging one for Case 2. In the NYC domain, the global average may not be an informative input to identify wetland temperature. The RSME is significantly reduced only when the proposed models receive more detailed inputs, as in Case 3. In contrast, in the PGH domain, wetland temperature shows the highest prediction error for all cases, except Baseline 1. The calibrated models might under-fit this small fraction (0.51%) of wetland, which still has significant heterogeneity.

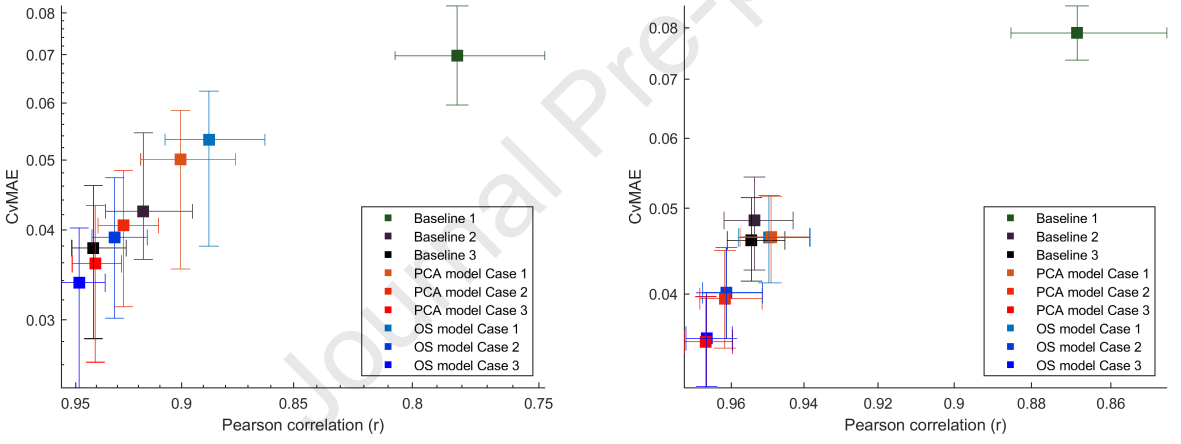
A further investigation on the model performance is conducted with b_{bias}, CvMAE, and correlation coefficient (r). Figure 9 and Figure 10 represent the statistical summary of the evaluated error metrics at individual locations. Figure 9 demonstrates that the proposed linear models produce less biased forecasts at overall locations and for the overall use cases. Figure 10 shows the statistical summary of CvMAE and correlation coefficient (r); the horizontal and the vertical range represent 25 and 75 percentiles with the mean values of the locations. For Case 3, the PCA model shows 0.036 and 0.035 CvMAE, having 0.94 and 0.97 r-values, in NYC and PGH, respectively. Meanwhile, the OS model shows 0.034-0.035 CvMAE and 0.95-0.97 r-values, for the same case. Baseline 3 slightly outperforms the PCA model in the NYC domain, but the difference is negligible. The assessment still confirms the ability of the proposed models as the forecasting systems.

Table 9 reports the evaluated RMSE, CRPS, and CE with additional test periods, which were adopted from [19]; the additional periods were set differently for the NYC and PGH domains. From Aug-01 to Aug-04 in 2019, the GP model reports its CRPS as 0.80°C and 0.72°C in the NYC and PGH domains, respectively, while CE is 2.1732 and 1.7222. In terms of CRPS, the proposed models show better performance for both domains during the period. In comparing CE values, the OS model outperforms the other two models (GP and PCA), while the other models show very similar performance. However, we argue that the PCA model is still more advantageous because of its computational efficiency, as shown in Table 10. In Table 9, the model performance is consistent for all three test periods, which hints at promising applications of the models for further model-use periods. Table 10 summarizes the measured computational times of the proposed models (Case 3) and Baseline 2 to generate the three days forecast. We choose Baseline 2 for the comparison since it is the fastest case among the baselines. Baselines 1 and 3 took more than an hour using the same



(a) New York City **(b)** Pittsburgh

Figure 9: Statistics summary for unnormalized bias (b_{bias}). In this figure, b_{bias} of each location is computed with pairs of simulated data and 24-hours ahead forecasts, then the box-plots report statistics of b_{bias} over the spatial domains.



(a) New York City **(b)** Pittsburgh

Figure 10: Statistics summary for error measures (r vs CvMAE). The error metrics of each location are evaluated with pairs of simulated data and 24-hours ahead forecast, then the sub-figures summarize statistics of the error metrics over the test regions. Each marker is centered at the mean of error metrics, providing 25 and 75 percentiles with their bounds. The axes for Pearson correlation (r) are plotted in descending order.

computing resource (12 processors of Intel Xeon E5-2690 v4@2.60GHz). As a result, the PCA and OS models provide qualified forecasts within a shorter evaluation time, 20-170sec, using one-twelfth of the computing resource (a single processor of Intel Xeon E5-2690 v4@2.60GHz). Such rapid computation has been achieved by the efficient state-space representation with the low dimensional embedding. The domain size and the different latent dimensions lead to the varying evaluation time of the proposed models. For example, the NYC domain (150km×150km) shows a shorter evaluation time than that of the PGH domain (200km×200km). Also, the latent dimension of the OS model (135 variables in NYC and 276 variables in PGH) is greater than that of the PCA model (46 variables in NYC and 91 variables in PGH), so the evaluation time is shorter with the PCA model.

The numerical campaign shows that the prediction performance is similar in PCA and OS models. However, from the authors' experience, the PCA model has a technical advantage in choosing the dimension of latent states. In the model calibration, the PCA model achieved a smaller reconstruction error with fewer latent variables than the OS model. The

Table 9
Probabilistic performance assessment for the test periods

Test period	Metric	New York City		Pittsburgh	
		PCA	OS	PCA	OS
Jul-15-2006 00:00	RMSE (°C)	1.03	1.00	-	-
~ Jul-21-2006 23:00 UTC	CRPS (°C)	0.51	0.48	-	-
	CE	2.60	1.66	-	-
Jun-15-2012 00:00	RMSE (°C)	-	-	1.06	1.02
~ Jun-21-2012 23:30 UTC	CRPS (°C)	-	-	0.59	0.56
	CE	-	-	2.02	1.48
Aug-01-2019 00:00	RMSE (°C)	1.13	1.06	0.99	0.97
~ Aug-04-2019 00:00 UTC	CRPS (°C)	0.61	0.55	0.55	0.53
	CE	2.20	1.50	1.72	1.38

Table 10
Evaluation time and computing resource for probabilistic forecast

Use case	Evaluation time (sec)		Computing resource
	New York City	Pittsburgh	
PCA model (case 3)	20.22	54.10	Intel Xeon E5-2690 v4@2.60GHz (1 processor)
OS model (case 3)	50.30	168.56	Intel Xeon E5-2690 v4@2.60GHz (1 processor)
Baseline 2	50.66	401.98	Intel Xeon E5-2690 v4@2.60GHz (12 processors)

rapidly decaying error of the PCA model enables the convenient dimension truncation and the faster computation in Table 10.

5. Conclusion

We proposed a probabilistic approach to provide a faster temperature forecast with reasonable accuracy as a potential alternative to physics-based numerical weather prediction models. The probabilistic spatio-temporal modeling has been developed to capture statistical temperature patterns and facilitate them to enable faster forecast. The method is based on linear Markov latent states of low dimensionality, alleviating the risk of overfitting, given the high dimensionality of the temperature data. The latent states account for the spatio-temporal evolution of the process. Also, such state-space representation substantially reduces computing time during the forecasting stage. In this paper, we presented two alternative linear models. The first model relies on global features, using principal component analysis, while the second one focuses on local measures of optimally sensing locations. Two models showed similar prediction performance, although technical preference to the PCA model may exist for its straightforward implementation.
We calibrated the proposed models with simulated temperature data using a physical simulator, the Princeton urban canopy model coupled with weather research and forecasting model, WRF-PUCM. For the data collection, the physics-based model was exploited to downscale historical weather data in the regions around New York City and Pittsburgh for three months of summer, June/July/August, from 2016 to 2018. Then, the Kalman Filter/Smoother scheme is implemented to integrate sensor data for the faster near-surface temperature forecast. We tested the model performance by forecasting a short-term near-surface temperature for August 1st to 3rd, 2019. Under various model case studies, the result confirms that the proposed models produce an adequate forecast at a relatively low computational cost. During the numerical campaign with the best use case, our models achieved 0.97-1.13°C root-mean-squared error for 24-hours ahead forecast, using only between 20 to 170 seconds to generate three days forecast with a single processor computer. The result is 14-22% improved accuracy compared to the most accurate use of the baseline model, which is based on a Gaussian process regression, and our models required less evaluation time than the fastest case of the baselines using only one twelve computing resource. This versatile modeling method is expected to expand its applications to further heat-induced risk analysis in urban areas. For example, having tested it here for one forecast realization, it can

be extended and applied to rapidly downscale the result of an ensemble of forecasts to better inform on the probabilistic heat risk for urban residents and allow decision makers to take early and appropriate action.

6. Data availability

The data presented in this paper are available through a Mendeley Data repository as follows:

- SHADE 2021: Meso-scale 2m air temperature around New York City and Pittsburgh
- Availability: https://data.mendeley.com/datasets/pr4m594vmf/1
- DOI: 10.17632/pr4m594vmf.1
- Size: 3.11 GB
- Year first available: 2021
- Email: byeongsc@andrew.cmu.edu.

Acknowledgement

This research is supported by the National Science Foundation under Grant no. 1664091 and 1664021. The authors thank Prof. Prathap Ramamurthy at The City College of New York for the assistance to simulate past climates with WRF-PUCM. Also, the project UPRI-0007 (Computational and Information Systems Laboratory) provided the computing resource for the research.

CRediT authorship contribution statement

Byeongseong Choi: Methodology, Software, Investigation, Formal analysis, Writing - Original Draft, Visualization, Data Curation. **Mario Berges:** Validation, Formal analysis, Writing - Review & Editing, Supervision. **Elie Bou-Zeid:** Software, Investigation, Resources, Writing - Review & Editing. **Matteo Pozzi:** Conceptualization, Validation, Formal analysis, Resources, Writing - Review & Editing, Supervision, Funding acquisition, Project administration.

References

[1] Anderson, B.G., Bell, M.L., 2009. Weather-related mortality: how heat, cold, and heat waves affect mortality in the united states. Epidemiology (Cambridge, Mass.) 20, 205. DOI: 10.1097/EDE.0b013e318190ee08.

[2] Research Data Archive at the National Center for Atmospheric Research, C., Laboratory, I.S., 2015. Ncep north american mesoscale (nam) 12 km analysis. DOI: 10.5065/G4RC-1N91.

[3] Barber, D., 2012. Bayesian reasoning and machine learning. Cambridge University Press.

[4] Berliner, L.M., Wikle, C.K., Cressie, N., 2000. Long-lead prediction of pacific ssts via bayesian dynamic modeling. Journal of climate 13, 3953–3968. DOI: 10.1175/1520-0442(2001)013<3953:LLPOPS>2.0.CO;2.

[5] Bird, A., Williams, C.K., 2019. Customizing sequence generation with multi-task dynamical systems. arXiv preprint arXiv:1910.05026 ArXiv:1910.05026.

[6] Changnon, S.A., Kunkel, K.E., Reinke, B.C., 1996. Impacts and responses to the 1995 heat wave: A call to action. Bulletin of the American Meteorological society 77, 1497–1506. DOI: 10.1175/1520-0477(1996)077<1497:IARTTH>2.0.CO;2.

[7] Choi, B., Berges, M., Bou-Zeid, E., Pozzi, M., 2021. Shade 2021: Meso-scale 2m air temperature around new york city and pittsburgh. DOI: 10.17632/pr4m594vmf.1.

[8] Chouzenoux, E., Elvira, V., 2020. Graphem: Em algorithm for blind kalman filtering under graphical sparsity constraints, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 5840–5844. DOI:10.1109/ICASSP40776.2020.9053646.

[9] Chow, W.T., Brennan, D., Brazel, A.J., 2012. Urban heat island research in phoenix, arizona: Theoretical contributions and policy applications. Bulletin of the American Meteorological Society 93, 517–530. DOI:10.1175/BAMS-D-11-00011.1.

[10] Cressie, N., Wikle, C.K., 2015. Statistics for spatio-temporal data. John Wiley & Sons. DOI: 10.1002/9781119115151.

[11] Gardner, W.A., Napolitano, A., Paura, L., 2006. Cyclostationarity: Half a century of research. Signal processing 86, 639–697. DOI:10.1016/j.sigpro.2005.06.016.

[12] Grossman-Clarke, S., Zehnder, J.A., Loridan, T., Grimmond, C.S.B., 2010. Contribution of land use changes to near-surface air temperatures during recent summer extreme heat events in the phoenix metropolitan area. Journal of Applied Meteorology and Climatology 49, 1649–1664. DOI:10.1175/2010JAMC2362.1.

[13] Guhathakurta, S., Gober, P., 2007. The impact of the phoenix urban heat island on residential water use. Journal of the American Planning Association 73, 317–329.
[14] Hannachi, A., Jolliffe, I.T., Stephenson, D.B., 2007. Empirical orthogonal functions and related techniques in atmospheric science: A review. International Journal of Climatology: A Journal of the Royal Meteorological Society 27, 1119–1152. DOI:10.1002/joc.1499.
[15] Krause, A., Singh, A., Guestrin, C., 2008. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. Journal of Machine Learning Research 9, 235–284.
[16] Lanzante, J.R., Dixon, K.W., Nath, M.J., Whitlock, C.E., Adams-Smith, D., 2018. Some pitfalls in statistical downscaling of future climate. Bulletin of the American Meteorological Society 99, 791–803. DOI: 10.1175/BAMS-D-17-0046.1.
[17] Li, D., Bou-Zeid, E., 2014. Quality and sensitivity of high-resolution numerical simulation of urban heat islands. Environmental Research Letters 9, 055001. DOI: 10.1088/1748-9326/9/5/05500.
[18] Li, D., Bou-Zeid, E., Barlage, M., Chen, F., Smith, J.A., 2013. Development and evaluation of a mosaic approach in the wrf-noah framework. Journal of Geophysical Research: Atmospheres 118, 11–918. DOI: 10.1002/2013JD020657.
[19] Malings, C., Pozzi, M., Klima, K., Bergés, M., Bou-Zeid, E., Ramamurthy, P., 2017. Surface heat assessment for developed environments: Probabilistic urban temperature modeling. Computers, Environment and Urban Systems 66, 53–64. DOI: 10.1016/j.compenvurbsys.2017.07.006.
[20] Malings, C., Pozzi, M., Klima, K., Bergés, M., Bou-Zeid, E., Ramamurthy, P., 2018. Surface heat assessment for developed environments: Optimizing urban temperature monitoring. Building and Environment 141, 143–154. DOI: 10.1016/j.buildenv.2018.05.059.
[21] Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S.P., Zimmerman, N., Kara, L.B., Presto, A.A., Subramanian, R., 2019. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. Atmospheric Measurement Techniques 12, 903–920. DOI: 10.5194/amt-12-903-2019.
[22] Maraun, D., Widmann, M., 2018. Statistical downscaling and bias correction for climate research. Cambridge University Press. DOI:10.1017/9781107588783.
[23] Min, Y.M., Kryjov, V.N., Oh, J.H., 2011. Probabilistic interpretation of regression-based downscaled seasonal ensemble predictions with the estimation of uncertainty. Journal of Geophysical Research: Atmospheres 116. DOI:10.1029/2010JD015284.
[24] Ramamurthy, P., Li, D., Bou-Zeid, E., 2017. High-resolution simulation of heatwave events in new york city. Theoretical and applied climatology 128, 89–102. DOI:10.1007/s00704-015-1703-8.
[25] Robinson, P.J., 2001. On the definition of a heat wave. Journal of applied Meteorology 40, 762–775. DOI: 10.1175/1520-0450(2001)040<0762:OTDOAH>2.0.CO;2.
[26] Rodrigues, E.R., Oliveira, I., Cunha, R., Netto, M., 2018. Deepdownscale: a deep learning strategy for high-resolution weather forecast, in: 2018 IEEE 14th International Conference on e-Science (e-Science), IEEE. pp. 415–422. DOI: 10.1109/eScience.2018.00130.
[27] Santamouris, M., Cartalis, C., Synnefa, A., Kolokotsa, D., 2015. On the impact of urban heat island and global warming on the power demand and electricity consumption of buildingsâĂŤa review. Energy and Buildings 98, 119–124. DOI: 10.1016/j.enbuild.2014.09.052.
[28] Särkkä, S., 2013. Bayesian filtering and smoothing. 3, Cambridge University Press.
[29] Shin, Y., Yi, C., 2019. Statistical downscaling of urban-scale air temperatures using an analog model output statistics technique. Atmosphere 10, 427. DOI: 10.3390/atmos10080427.
[30] Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Liu, Z., Berner, J., Wang, W., Powers, J.G., Duda, M.G., Barker, D.M., et al., 2019. A description of the advanced research wrf model version 4. National Center for Atmospheric Research: Boulder, CO, USA , 145.
[31] Trevor, H.Z., Zou, H., Hastie, T., Tibshirani, R., 2004. Sparse principal component analysis, in: Journal of Computational and Graphical Statistics, Citeseer.
[32] Voogt, J.A., Oke, T.R., 2003. Thermal remote sensing of urban climates. Remote sensing of environment 86, 370–384. DOI: 10.1016/S0034-4257(03)00079-8.
[33] Wallemacq, P., House, R., 2018. Economic losses, poverty, and disasters: 1998-2017.
[34] Wang, Z.H., Bou-Zeid, E., Smith, J.A., 2013. A coupled energy transport and hydrological model for urban canopies evaluated using a wireless sensor network. Quarterly Journal of the Royal Meteorological Society 139, 1643–1657. DOI: 10.1002/qj.2032.
[35] Welch, G., Bishop, G., et al., 1995. An introduction to the Kalman filter. Technical Report. Department of Computer Science, University of North Carolina at Chapel Hill.
[36] Weyn, J.A., Durran, D.R., Caruana, R., 2019. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. Journal of Advances in Modeling Earth Systems 11, 2680–2693. DOI: 10.1029/2019MS001705.
[37] Wilby, R., Dawson, C., Barrow, E., 2002. sdsm âĂŤ a decision support tool for the assessment of regional climate change impacts. Environmental Modelling & Software 17, 145–157. DOI:10.1016/S1364-8152(01)00060-3.
[38] Zuo, J., Pullen, S., Palmer, J., Bennetts, H., Chileshe, N., Ma, T., 2015. Impacts of heat waves and corresponding measures: a review. Journal of Cleaner Production 92, 1–12. DOI: 10.1080/01944360708977980.

A. Appendix A

In this section, we discuss the relationship between the calibrating procedure, proposed in this paper, and the iterative expectation maximization (EM) algorithm. With Bayes formula, the posterior distribution $p(\mathbf{x}|\mathbf{y})$ can be decomposed into the prior distribution $p(\mathbf{x})$ and the likelihood $p(\mathbf{y}|\mathbf{x})$ as the following:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{A.1}$$

For multidimensional Gaussian random variables $\mathbf{x}$ and $\mathbf{y}$, the posterior distribution is also known to follow Gaussian distribution, and the following formula describes the posterior mean as a linear combination of the prior mean and its corresponding likelihood estimator.

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}{\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1}\boldsymbol{\mu}_{\pi} + \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}{\boldsymbol{\Sigma}_{\mathbf{L}}}^{-1}\boldsymbol{\mu}_{L} \tag{A.2}$$

$$\begin{aligned} \text{where} \quad \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} &= \left[{\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1} + {\boldsymbol{\Sigma}_{\mathbf{L}}}^{-1}\right]^{-1} \\ \text{or} \quad \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} &= \boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}\mathbf{x}} \end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}$ are the prior covariance of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\boldsymbol{\Sigma}_{\mathbf{L}}$ is the covariance of $\mathbf{x}$ that is evaluated from the likelihood function with the given observation $\mathbf{y}$. $\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}$ is the posterior covariance matrix of $\mathbf{x}$, conditional to $\mathbf{y}$. Now, we consider the graphical representation of the zero-mean state-space model as Figure A.1a. In Figure A.1b, the distribution of $\mathbf{x}_t$ is set as a prior distribution with the input states $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t+1}$ without inclusion of the observation on $\mathbf{y}_t$ (red dotted box); here, the prior distribution is to denote it is prior to observe the emission $\mathbf{y}_t$. Adopting the notations of [3], the prior mean of $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t+1})$ can be represented as the following due to the linearity between Gaussian random variables.

$$\boldsymbol{\mu}_{\pi} = \mathbf{A}_{\pi}\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{x}_{t-1}} + \overleftarrow{\mathbf{A}}_{\pi}\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{x}_{t+1}} \tag{A.3}$$

$$\begin{aligned} \text{where} \quad \mathbf{A}_{\pi} &= \boldsymbol{\Sigma}_{\pi}\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_t|\mathbf{x}_{t-1}}, \\ \overleftarrow{\mathbf{A}}_{\pi} &= \boldsymbol{\Sigma}_{\pi}\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_t|\mathbf{x}_{t+1}} \\ \boldsymbol{\Sigma}_{\pi} &= \left[\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_t|\mathbf{x}_{t-1}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{x}_t|\mathbf{x}_{t+1}}\right]^{-1} \end{aligned}$$

$\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{x}_{t-1}}$ and $\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{x}_{t+1}}$ denotes the conditional mean of $\mathbf{x}_t$ by the dynamic predictors $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t+1}$, respectively; similarly, $\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{x}_{t-1}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{x}_{t+1}}$ are the conditional covariance by the predictors. Now, we update the prior distribution again with the observation $\mathbf{y}_t$, which is far high dimensional variable than the latent $\mathbf{x}_t$ (Figure A.1c and Figure A.1d).

$$\boldsymbol{\mu}_{\text{post}} = \mathbf{B}_{\pi}\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{x}_{t-1}} + \overleftarrow{\mathbf{B}}_{\pi}\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{x}_{t+1}} + \mathbf{B}_{L}\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{y}_t} \tag{A.4}$$

$$\begin{aligned} \text{where} \quad \mathbf{B}_{\pi} &= \boldsymbol{\Sigma}_{\text{post}}\boldsymbol{\Sigma}_{\pi}^{-1}\mathbf{A}_{\pi}, \\ \overleftarrow{\mathbf{B}}_{\pi} &= \boldsymbol{\Sigma}_{\text{post}}\boldsymbol{\Sigma}_{\pi}^{-1}\overleftarrow{\mathbf{A}}_{\pi} \\ \mathbf{B}_{L} &= \boldsymbol{\Sigma}_{\text{post}}\boldsymbol{\Sigma}_{L}^{-1} \\ \boldsymbol{\Sigma}_{\text{post}} &= \left[\boldsymbol{\Sigma}_{\pi}^{-1} + \boldsymbol{\Sigma}_{L}^{-1}\right]^{-1} \end{aligned}$$

$\boldsymbol{\mu}_{\mathbf{x}_t|\mathbf{y}_t}$ and $\boldsymbol{\Sigma}_{L}$ represent the likelihood estimators for the mean and the covariance of $\mathbf{x}_t$ with the observed emission $\mathbf{y}_t$. As a forward and backward smoothing scheme, EM-algorithm iteratively improve the state variable $\mathbf{x}_t$ and the transition matrices $\mathbf{A}_{\pi}$ and $\overleftarrow{\mathbf{A}}_{\pi}$. As the analysis on non-EM implementation, we compare the relative magnitude of the forward weight $\mathbf{B}_{\pi}$ and the backward weight $\overleftarrow{\mathbf{B}}_{\pi}$ to the likelihood weight $\mathbf{B}_{L}$, using the calibrated parameters in 3. The defined relative magnitude is the trace of forward or backward weight, being normalized by that of the likelihood weight. Figure A.2 summarizes the result of the numerical test presenting the relative contributions of the dynamic predictors. According to the result, the contributions of forward and backward priors are relatively small to that of the likelihood, so the likelihood shows the dominant impact on the states estimation. Therefore, the forward and backward smoothing scheme, in EM algorithm, is not expected to yield much improvement.

A.1. Example: City of Pittsburgh

Implementing EM algorithm for high dimensional system consumes additional computing resource for the model calibration. In this subsection, we continue the discussion on EM versus non-EM implementation with a reduced-scale

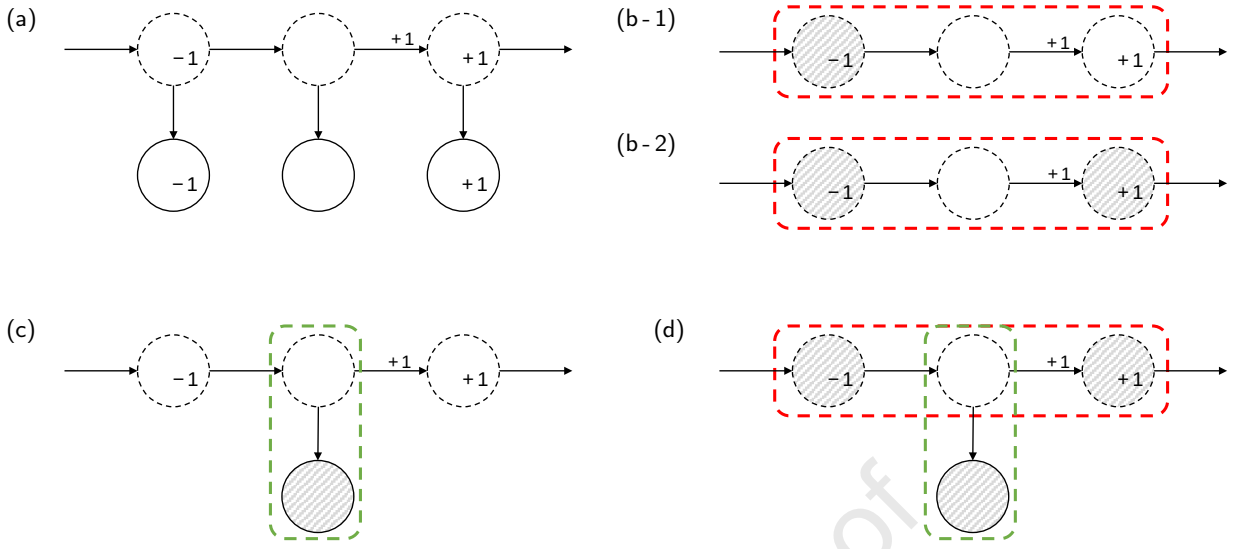
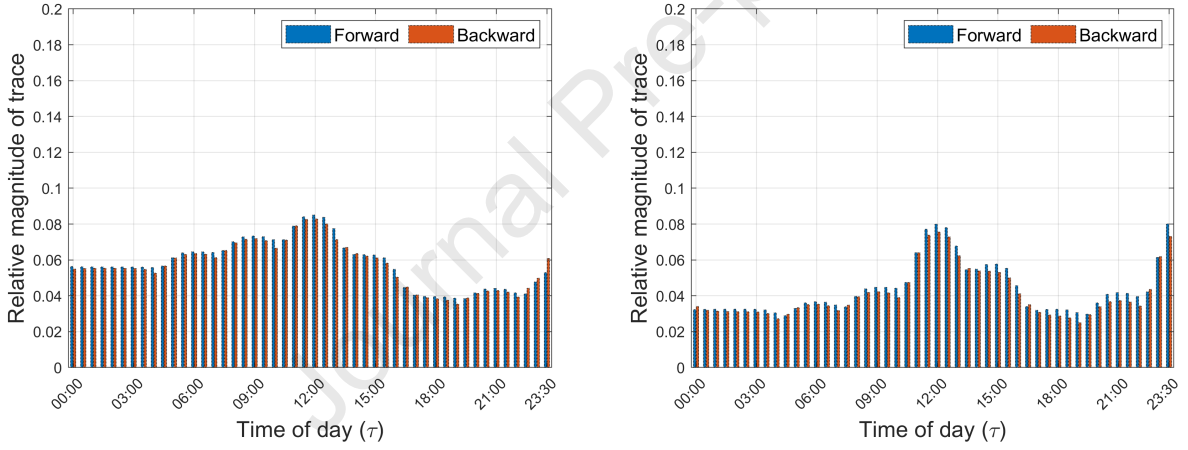


Figure A.1: Graphical representation of a state-space model



(a) New York **(b)** Pittsburgh

Figure A.2: Relative magnitude of prior weights

model. Figure A.3 illustrates the map of the model at City of Pittsburgh. The mass points represent the locations of the interest within the city. Using EM algorithm, we calibrate the PCA model, then show the change by the algorithm. For this city-scaled model, only 3 principal components are enough to achieve the 0.35°C and 0.36°C reconstruction error for EM and non-EM, respectively. We calculate the likelihood function for each single EM step, dividing the term into transition (prior) and emission (likelihood) parts. Figure A.4 shows the convergence of the EM algorithm, and we scattered the estimated latent variables at the last stage of the iterative algorithm with those of non-EM implementation (Figure A.5). As the algorithm progresses, the total likelihood increases with small increments, but the trade-off exists for the transition and the emission. The numerical investigation confirms that the improvement is not significant, especially for the lower-order principal components, which account for dominant behavior of the system. Therefore, the direct measurement of the emission still has the most significant contribution to the latent states inference.

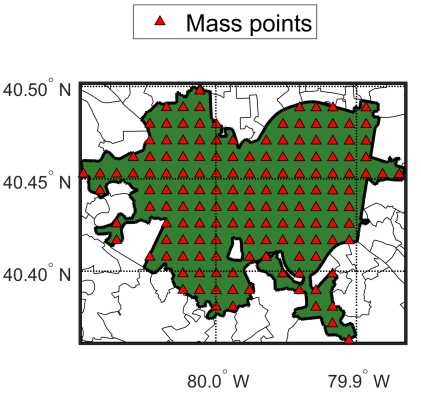


Figure A.3: Mass points within Pittsburgh

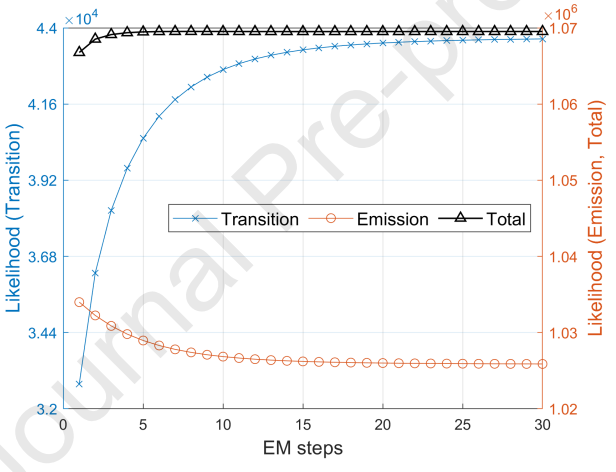


Figure A.4: Convergence diagram

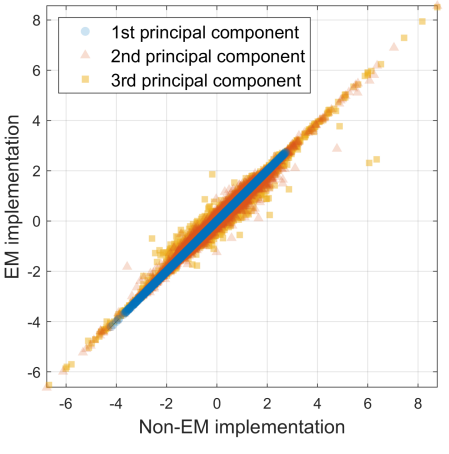


Figure A.5: Estimated latent variables

Highlights

Developing short-term probabilistic forecasts of meso-scale near-surface urban temperature fields

Byeongseong Choi,Mario Berges,Elie Bou-Zeid,Matteo Pozzi

- We develop probabilistic models for meso-scale near-surface urban air temperature.
- We calibrate/validate the models on simulated data for New York City and Pittsburgh.
- A Kalman filter/smoother updates the proposed models for adaptive forecast.
- The proposed models use 3 to 8% the computing resources used by a comparable model.
- 24-hours ahead forecasts show 0.97-1.13°C prediction error.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mario Berges reports financial support was provided by National Science Foundation. Elie Bou-Zeid reports financial support was provided by National Science Foundation. Matteo Pozzi reports financial support was provided by National Science Foundation.