Bias correction in daily maximum and minimum temperature measurements through Gaussian process modeling

Maxime Rischard[*], Karen A. McKinnon[**], and Natesh Pillai[*]

[*]Department of Statistics, Harvard University

[**]National Center for Atmospheric Research; Descartes Labs

April 18, 2018

Abstract

The Global Historical Climatology Network-Daily database contains, among other variables, daily maximum and minimum temperatures from weather stations around the globe. It is long known that climatological summary statistics based on daily temperature minima and maxima will not be accurate, if the bias due to the time at which the observations were collected is not accounted for. Despite some previous work, to our knowledge, there does not exist a satisfactory solution to this important problem. In this paper, we carefully detail the problem and develop a novel approach to address it. Our idea is to impute the hourly temperatures at the location of the measurements by borrowing information from the nearby stations that record hourly temperatures, which then can be used to create accurate summaries of temperature extremes. The key difficulty is that these imputations of the temperature curves must satisfy the constraint of falling between the observed daily minima and maxima, and attaining those values at least once in a twenty-four hour period. We develop a spatiotemporal Gaussian process model for imputing the hourly measurements from the nearby stations, and then develop a novel, easy to implement, MCMC technique to sample from the posterior distribution satisfying the above constraints. We validate our imputation model using hourly temperature data from four meteorological stations in Iowa, of which one is hidden and the data replaced with daily minima and maxima, and show that the imputed temperatures closely match the hidden temperatures. We also demonstrate that our model can exploit information contained in the data to infer the time of daily measurements.

1 Introduction

Long, high-quality records of temperature provide an important basis for our understanding of climate variability and change. Historically, there has been a focus on monthly-average temperature records that are sufficient for certain analyses, such as quantifying long-term changes in temperature. As our knowledge of climate change expands, however, there is increasing interest in understanding changes in temperature on shorter timescales, with a particular focus on extreme events. To do so, it is necessary to utilize temperature data with higher temporal resolution.

Recent work has led to the development of the Global Historical Climatology Network-Daily (GHCND) database [8], which contains, among other variables, daily maximum and minimum temperatures from weather stations around the globe. The database draws from a range of different sources, and the data within it undergoes basic quality control to remove erroneous values.

The current quality control methodology, however, does not account for so-called 'inhomogeneities'. Inhomogeneities result from changes in measurement practices that impact the recorded temperatures. For temperature, known inhomogeneities include (a) changes in the time of observation, (b) changes in the thermometer technology, (c) station relocation, and (d) changes in land use around a station [10]. While these inhomogeneities have a small effect on, for example, the estimation of global mean temperature, they can have a large effect on estimation of temperature variability and change at a more local scale.

There is a large body of work focused on homogenizing monthly-average temperatures [e.g., 7, 6, 11, 5, 9, 16], resulting in widely available, large-scale homogenized monthly temperature datasets. Homogenization typically proceeds through identifying non-climatic 'breakpoints' in a given time series through comparison with neighboring stations. Once a breakpoint is identified, the measurements recorded after the breakpoint are adjusted in some way to reduce or remove the inhomogeneity. Most applications of these methods, however, focus on adjusting the mean state of the data rather than the shape of the distribution [see 4, and references therein]. While this may be sufficient for monthly data, it is known that changes in measurement practices may affect different quantiles of the daily temperature distribution unequally. To address this issue, some homogenization methods have also employed frequency distribution matching techniques, so that each temperature recorded after a breakpoint is adjusted according to its percentile within the time series (percentile of what?) [4, 15].

1.1 The problem

Many historical measurements of daily temperatures are provided as daily maxima and minima, which should bracket the diurnal cycle for each day. In this section we illustrate the bias in the measurements of

maximum and minimum temperatures (T_x and T_n respectively) for a given cycle due to the time of day at which observations are recorded. This bias exists because T_x and T_n are often recorded by an observer who visits a weather station every 24 hours, and notes the maximum and minimum temperatures measured by the thermometer over the previous 24 hours.

Ideally, the observer would visit the station at midnight, and the recorded highest and lowest temperatures over the past 24 hours would be representative of the maximum and minimum of the diurnal cycle of the previous day.

Figure 1 illustrates this bias in daily maxima and minima with ten days of hourly temperature measurements from the Waterloo Municipal Airport station in Iowa. Ideally, T_x measurements should capture the peak of each diurnal cycle, and T_n its trough. We emulate daily T_x/T_n measurements by dividing the data into 24 hour measurement windows, and reporting the minimum and maximum temperature that was recorded in this window. On most days, the measurements capture the peak and trough of the diurnal cycle: the triangles and lines coincide. But there are also several discrepancies, typically in T_x when the measurements are made near the warmest hour of the day, and in T_n when the measurements are made near the coldest hour. The most blatant example occurs on April 3rd, where the peak of the diurnal cycle is 7.2°C and occurs at 21:00 (all times are in the UTC-6 time zone, and tick marks are at midnight at the start of each day), but the day's T_x record of 16.1°C is reached immediately after the previous day's measurement. The measured T_x therefore overestimates the diurnal cycle's peak by 8.9°C. Ideally, measurements of the diurnal cycle peak and trough would be obtained by recording T_x and T_n at the coldest and warmest time of day respectively. This choice minimizes the possibility of the previous or next diurnal cycle affecting the measurement. For convenience, however, most observers record data at a single daytime hour instead. Consequently, measurements recorded in the early morning may not properly register the low of the previous night if it was unusually warm. Similarly, measurements recorded in the afternoon may not properly register the afternoon's peak temperature if it was usually cool. Our goal is to address the bias that results from this measurement practice.

The bias in the daily records can in turn induce bias in the long-term summary statistics that are of climatological interest. A measure as simple as the average daily maximum temperature for an entire year (2015) increases by over 1°C if the measurements are made at 15:00 compared to 9:00 (see Figure 2). Conversely, the average T_n is colder by over 1°C if T_n is measured at 5:00 (the coldest time of day on average) rather than 15:00.

If the time of observation remained constant over time, this systematic bias would still exist, but it would not be linked to spurious trends in the data. However, there have been known (and likely unknown) changes in the time of observation. In the United States, for example, observers were instructed to switch

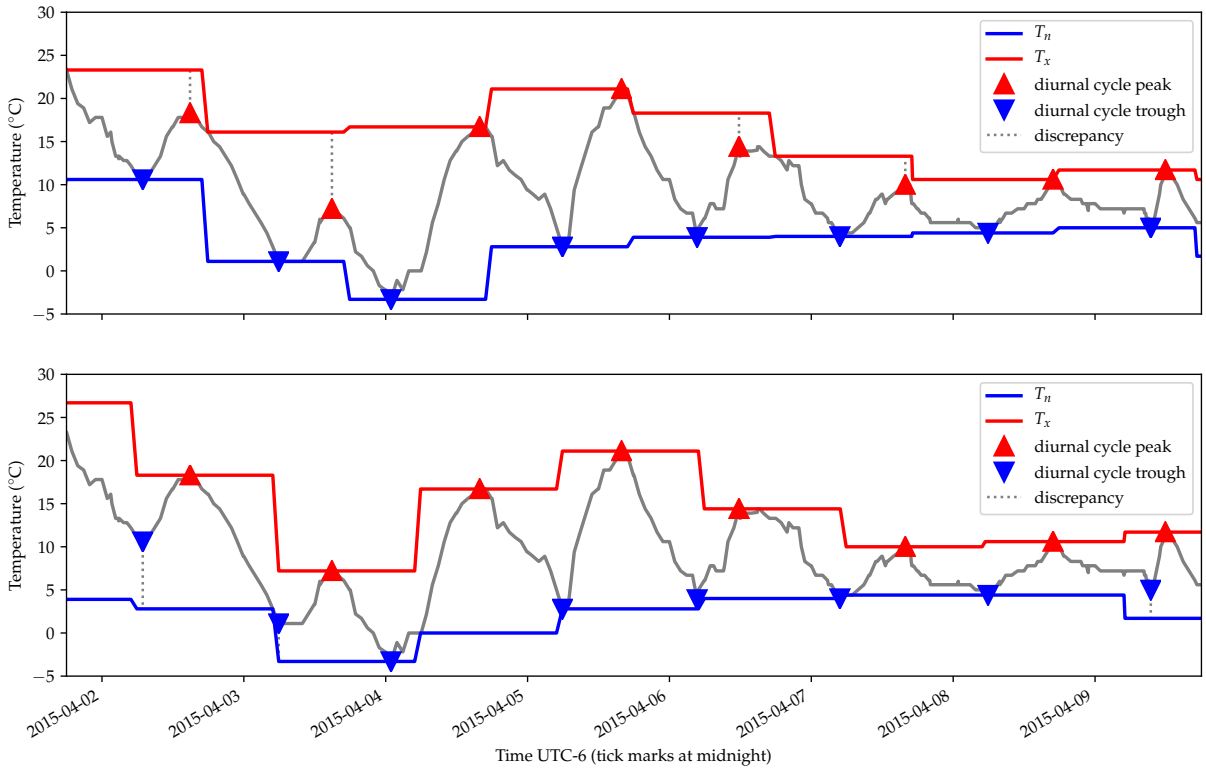

Figure 1: *An extract of the temperature measurements made at the Waterloo Municipal Airport. The blue and red triangles respectively indicate the coldest and warmest temperature of each diurnal cycle. The blue and red lines respectively denote the observed maximum and temperature recorded each day at 17:00 (top) or 5:00 (bottom) for the 24-hour period preceding the measurement. The occasional discrepancies between the 24-hour extrema, and the peaks and troughs of the diurnal cycle, are indicated with dotted lines.*

from recording data in the afternoon to recording data in the morning beginning in the 1950s. This change has led to an apparent decrease in both T_x and T_n over time [10].

A climatologist studying weather variability might be interested in summary statistics such as the average absolute change in the daily temperature maxima and minima from one day to the next. The answer to this question also depends on the time of day at which the temperatures are recorded. Collecting the measurements at the hottest time of day means that the peaks on a warm day gets recorded twice, erasing the diurnal peaks of the following colder day, and hence the variability gets underestimated. We can see this in Figure 3, where the respective variability estimates drop if the maxima get measured at the warmest time, or if the minima get measured at the coldest time.

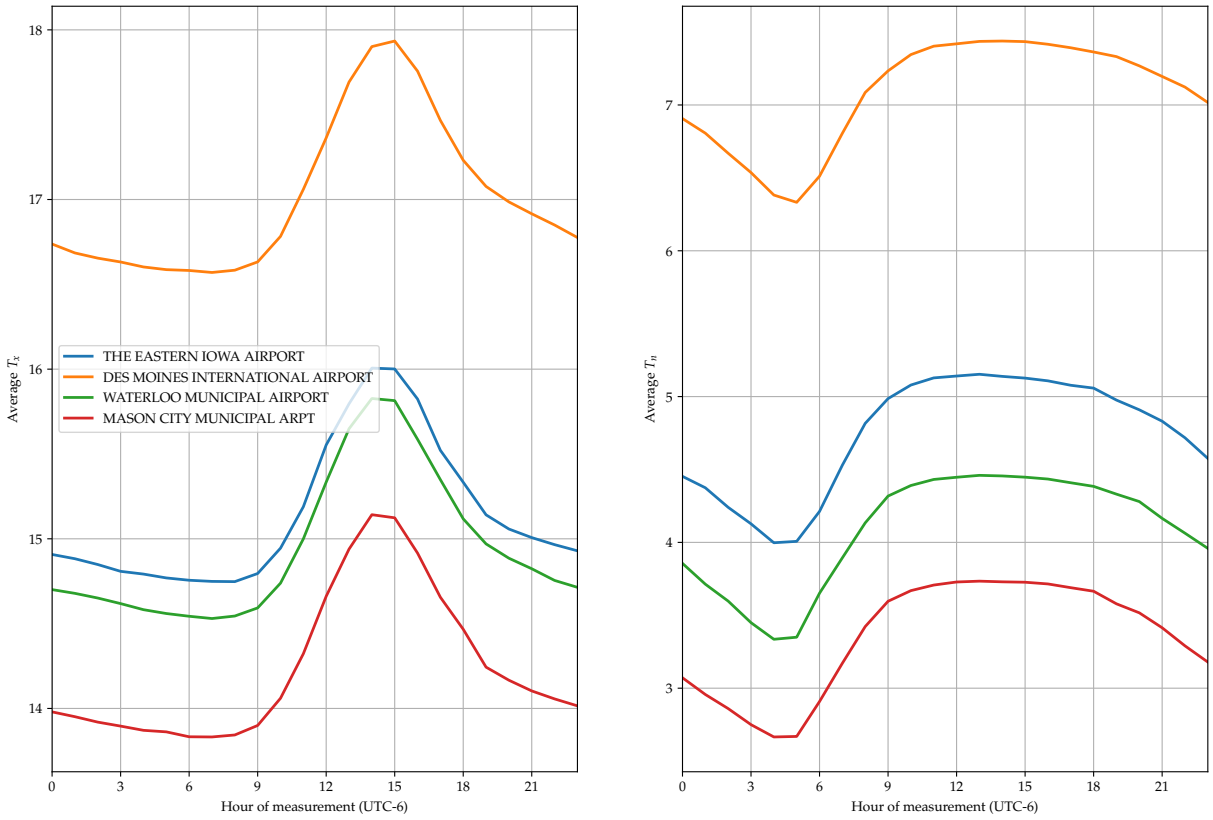

Figure 2: *Mean daily maximum temperature (left) and mean daily minimum temperature (right) extracted from hourly measurements at four Iowa weather stations in 2015 under varying measurement times (UTC-6) .*

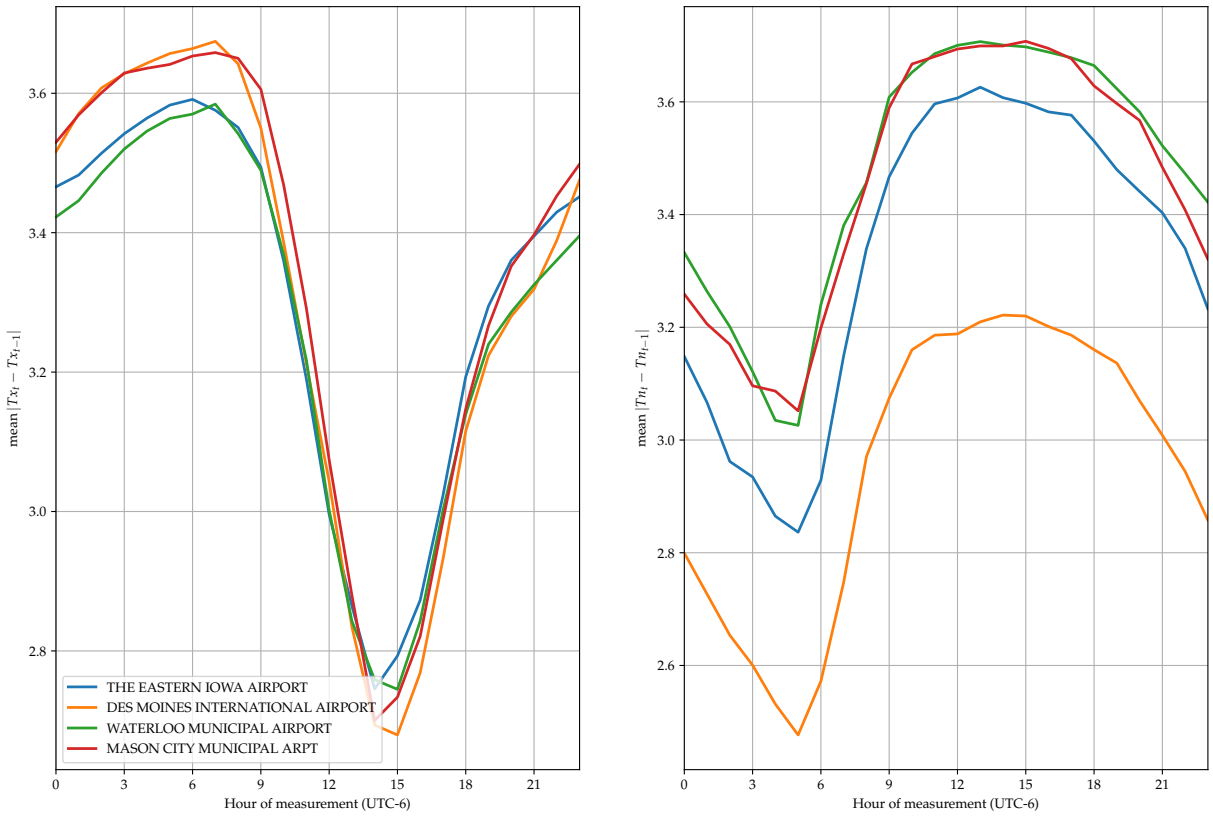

Figure 3: *Mean absolute daily change in maximum temperature (left) and mean absolute daily change in minimum temperature (right) extracted from hourly measurements at four Iowa weather stations in 2015 under varying measurement times (UTC-6).*

1.2 Our approach and proposed solution

The goal of our approach is to infer the true T_x and T_n values throughout the data records, thereby correcting both the variance biases and the spurious trends. This stands in constrast to previous work, which has focused primarily on addressing spurious trends. We approach the problem as a missing data problem, wherein we seek to recover the values of T_x and T_n that may have been overwritten due to measurement practices. Our idea is to impute the hourly temperatures at the location of the measurements by borrowing information from the nearby stations that record hourly temperatures, which then can be used to create accurate summaries of temperature extremes. These hourly measurements at neighboring weather stations are considered less reliable by climatologists, as they are not as carefully documented, calibrated, and situated. For instance, weather stations at locations experiencing a lot of human activity, like airports, may record higher daily temperatures on average. Therefore, summary statistics extracted directly from those measurements would not be directly usable for climatology, as they could suffer from systematic bias. However, even if mis-calibrated, the meteorological data from these nearby stations do contain valuable information about the hourly changes in temperatures on any given day.

In this paper, we develop a spatiotemporal Gaussian process model using the information from nearby stations with hourly data and simulate multiple realizations of hourly temperature time series at each station of interest. The key technical difficulty is that these imputations of the temperature curves must satisfy the constraint of falling between the observed daily minima and maxima, and attaining those values at least once in a twenty-four hour period. We develop a novel, easy to implement, MCMC algorithm to sample from the posterior distribution satisfying the above constraints. Our constrained imputations are implemented in the Stan programming language [3]; our code is publicly available on the first author's GitHub account.

2 A First Spatiotemporal Model

To model measured temperatures at various locations and times, we develop a spatio-temporal Gaussian process model. In its simplest form, we posit that temperatures from stations that are near each other are more correlated than distant stations, and that those correlations should also decay in time. Squared exponential covariance functions can be used to model correlation decaying as a function of distance. We model the simultaneous temperatures throughout a region as a Gaussian process, with covariance between two locations $\mathbf{x}$ and $\mathbf{x}'$ given by the squared exponential kernel with characteristic lengthscale ℓ_x and variance

σ^2_{GP}:

$$\operatorname{cov}\left(T(\mathbf{x}), T(\mathbf{x}') \mid t\right) = k_{space}(\mathbf{x}, \mathbf{x}') = \sigma^2_{space} \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^\intercal (\mathbf{x}-\mathbf{x}')}{2\ell_x^2}\right). \tag{1}$$

Similarly, the time series of temperatures at a single location can be modeled as a Gaussian process with characteristic timescale ℓ_t:

$$\operatorname{cov}\left(T(t), T(t') \mid \mathbf{x}\right) = k_{time}(t, t') = \sigma^2_{time} \exp\left(-\frac{(t-t')^2}{2\ell_t^2}\right). \tag{2}$$

We then combine the spatial and temporal model by multiplying the correlation functions:

$$k_{st}(\mathbf{x}, \mathbf{x}', t, t') = k_{time}(t, t') \cdot k_{space}(\mathbf{x}, \mathbf{x}'). \tag{3}$$

This yields the covariance of the Gaussian process underlying the spatio-temporal model of temperatures. The variances σ^2_{space} and σ^2_{time} are not separately identifiable, so we arbitrarily fix $\sigma^2_{space} = 1$. To allow for systematic differences between stations, we add a mean temperature parameter $\mu_{station[i]}$ for each station, where station $[i]$ is the index of the station at which observation i was recorded. This parameter captures both systematic differences in temperature between locations, for example due to differences in altitude, vegetation, or built environment around the station, and also calibration errors in the measurement apparatus.

The observation model depends on the type of measurement obtained at a given location. At stations j that provide a full temperature time series, we model the i^{th} temperature record as a noisy measurement from the true time series, with iid normal noise:

$$\begin{aligned} T_{ij} &= \mu_j + f(\mathbf{x}_j, t_{ij}) + \epsilon_{ij}, \\ f(\mathbf{x}_j, t_{ij}) &\sim \mathcal{GP}\left(0, k_{st}(\mathbf{x}, \mathbf{x}', t, t')\right), \\ \epsilon_{ij} &\overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_\epsilon\right). \end{aligned} \tag{4}$$

The noise term captures measurement error, including rounding errors, and temperature fluctuations occuring on a very short time scale. At stations that only provide daily T_x and T_n records, we denote the time of the d^{th} daily measurement by t^{meas}_d, and approximate the T_x and T_n observation respectively as the maximum or minimum temperatures at a discretized set of times inside of $(t^{meas}_{d-1}, t^{meas}_d]$:

$$\begin{aligned} (T_x)_{dj} &= \max\left\{T_{ij}, \text{for all } i \text{ such that } t_{ij} \in (t^{meas}_{d-1}, t^{meas}_d]\right\}, \\ (T_n)_{dj} &= \min\left\{T_{ij}, \text{for all } i \text{ such that } t_{ij} \in (t^{meas}_{d-1}, t^{meas}_d]\right\}, \end{aligned} \tag{5}$$

with T_{ij} modeled as in (4).

2.1 Fitting the spatiotemporal model

Software is readily available in many programming languages for fitting Gaussian process models, including inference on the covariance parameters. We chose to use the julia `GaussianProcesses.jl` package to fit the above spatiotemporal model to the hourly temperatures at four Iowa weather stations. The Iowa data set includes 47,864 measurements, which is computationally challenging to fit directly with a single Gaussian process. There are many methods to handle large data sets with Gaussian processes: for example [12] review sparse approximations to Gaussian processes from a machine learning perspective, while [1] develop a method specifically for large spatial data sets. For simplicity, we chose instead to divide the data into 10-day chunks, modeled as independent Gaussian processes with shared hyperparameters. We put weak normal priors on $\mu_{\text{station}[i]}$ with large standard deviation $\sigma_\mu = 10\,^\circ$ C, which can be incorporated into the Gaussian process with an additional term

$$k_\mu(\mathbf{x}, \mathbf{x}') = \begin{cases} \sigma_\mu^2 & \text{if } \mathbf{x} = \mathbf{x}', \\ 0 & \text{otherwise}. \end{cases} \tag{6}$$

added to the covariance function. The spatio-temporal covariance function becomes

$$k_{st}(\mathbf{x}, \mathbf{x}', t, t') = k_{time}(t, t') \cdot k_{space}(\mathbf{x}, \mathbf{x}') + k_\mu(\mathbf{x}, \mathbf{x}'). \tag{7}$$

Our model thus has 4 parameters, σ_{GP}, ℓ_t, ℓ_x and σ_ϵ, which we fit by maximizing the marginal likelihood of the Iowa data

$$\widehat{\sigma_{GP}}, \widehat{\ell_t}, \widehat{\ell_x}, \widehat{\sigma_\epsilon} = \underset{\sigma_{GP}, \ell_t, \ell_x, \sigma_\epsilon}{\arg\max} \left\{ \mathbb{P}\left(T \mid \sigma_{GP}, \ell_t, \ell_x, \sigma_\epsilon\right) \right\}. \tag{8}$$

We obtained $\widehat{\sigma_{GP}} = 3.73\,^\circ$ C, $\widehat{\ell_t} = 2.7$ hours, $\widehat{\ell_x} = 176.4$ km and $\widehat{\sigma_\epsilon} = 0.44\,^\circ$ C.

3 Predictions using nearby data

After fitting our spatio-temporal Gaussian process model (4) with optimized covariance parameters, we use it to generate predictions at the station where we aim to generate imputations based on nearby measurements. Gaussian processes yield closed-form expressions for the posterior distribution of the imputed temperatures. We will denote the temperatures we wish to impute as T_{miss} at times t_{miss} and location $\mathbf{x}_{\text{miss}}$ and those observed at nearby stations as T_{nearby}, at times t_{nearby} and locations X_{nearby}. Under the spatio-

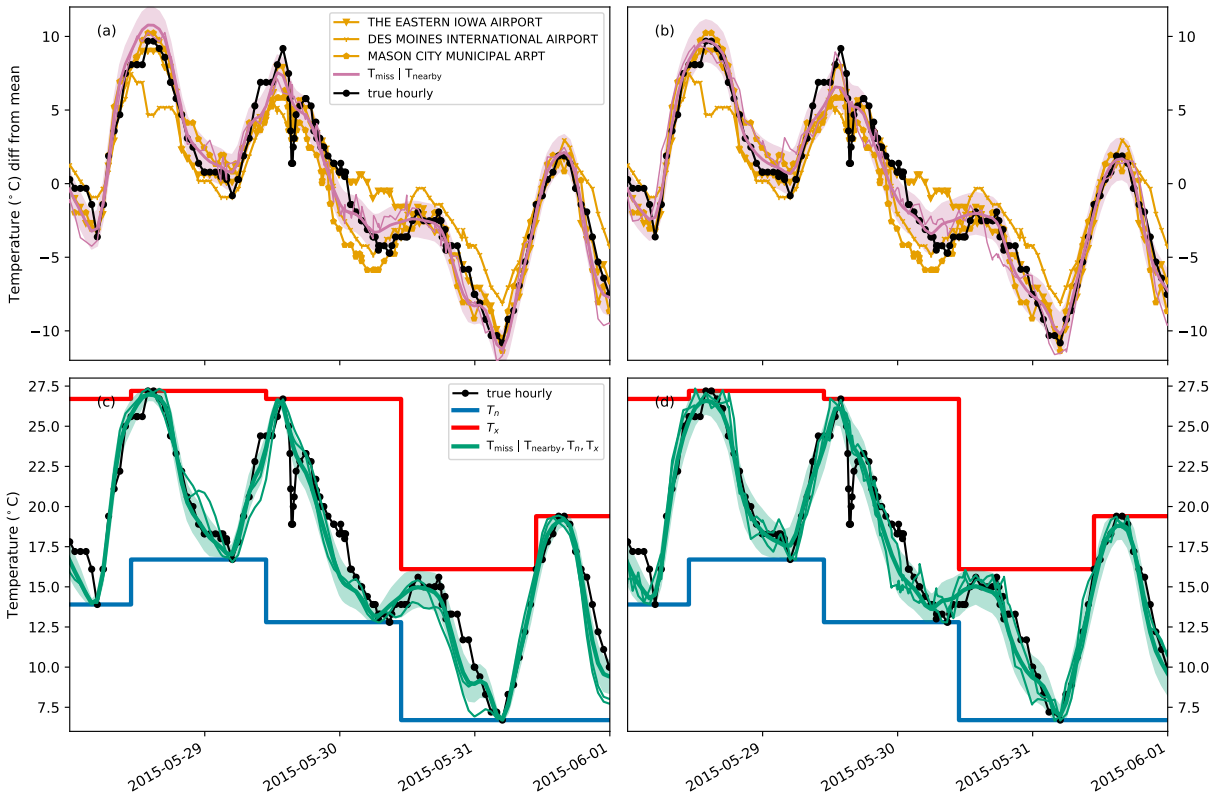

Figure 4: *Imputations of the temperature time series at Waterloo Municipal Airport between May 28, 2015 and June 1, 2015 (a) using only nearby data and the product of squared exponentials model; (b) using only nearby data and the sum of products model; (c) incorporating* $\mathrm{T_n}$ *and* $\mathrm{T_x}$ *measurements under the product of squared exponentials model; and (d) incorporating* $\mathrm{T_n}$ *and* $\mathrm{T_x}$ *measurements under the sum of products model. The mean is subtracted from each time series in (a) and (b) as the models leave the average temperature at the imputation site as a free parameter. For each imputation distribution, the mean is shown as a thick line, surrounded by an 80% credible envelope in lighter color, and example imputations as thinner lines.*

temporal model (4), T_{miss} and T_{nearby} are jointly multivariate normal, with mean zero and covariance given by $k_{st}(x, x', t, t')$. Standard results for conditioning within multivariate normals then yield

$$\begin{aligned} T_{miss} \mid T_{nearby} &\sim \mathcal{N}\left(\mu_{miss|nearby}, \Sigma_{miss|nearby}\right) , \\ \mu_{miss|nearby} &= \mathbb{E}\left(T_{miss} \mid T_{nearby}\right) \\ &= \operatorname{cov}\left(T_{miss}, T_{nearby}\right) \operatorname{cov}\left(T_{nearby}, T_{nearby}\right)^{-1} T_{nearby} , \\ \Sigma_{miss|nearby} &= \operatorname{var}\left(T_{miss} \mid T_{nearby}\right) \\ &= \operatorname{cov}\left(T_{miss}, T_{miss}\right) - \operatorname{cov}\left(T_{miss}, T_{nearby}\right) \operatorname{cov}\left(T_{nearby}, T_{nearby}\right)^{-1} \operatorname{cov}\left(T_{nearby}, T_{miss}\right) . \end{aligned} \tag{9}$$

All covariance matrices can be obtained by plugging into k_{st}. For example, the ij^{th} entry of $\operatorname{cov}\left(T_{miss}, T_{nearby}\right)$ is given by $k_{st}(x_{miss}, X_{nearby}[j], t_{miss}[i], t_{nearby}[j])$, where $X_{nearby}[j]$ gives the spatial covariates of the jth observation, and $t_{nearby}[j]$ its time.

In Figure 4(a), we show an example of predictions obtained from this spatio-temporal model. We withheld measurements from the Waterloo Municipal Airport, and then used data from three nearby stations between May 2, 2015 and May 5, 2015 to predict the Waterloo temperatures during the same time window. This allows us to assess the quality of the predictions on this example.

4 Imputations

4.1 Imputing by Conditioning on Extrema

Our aim is not simply to predict temperatures at a location with no measurements, but rather to impute hourly temperatures at a location with accurate measurements of the daily temperature extrema. This is an instance of a more general statistical problem: if a random p-vector $\{X_i : i = 1, \ldots, p\}$ has a known distribution F_X, and its maximum $X_{max} \equiv \max_i \{X_i\}$ and minimum $X_{min} \equiv \min_i \{X_i\}$ are measured, how does one draw samples from $F_{X|X_{max},X_{min}}$, the distribution of X conditional on X_{max} and X_{min}? Conditional draws from $F_{X|X_{max},X_{min}}$ need to respect three constraints: one component of X must be equal to X_{min}, another to X_{max}, and all other components must lie between X_{min} and X_{max}.

Conceptually, we could implement a valid imputation algorithm by drawing random samples F_X, and accepting only those samples that satisfy the three constraints. Unfortunately, if F_X is a continuous distribution, the probability of a random draw from F_X satisfying such sharp constraints is zero. One could envision adding some tolerance, so that samples with minimum and maximum within a small margin of

X_{max} and X_{min} are retained, but as the dimensionality p grows, the rejection probability will rapidly go to 1, thus requiring huge sample sizes. Ultimately, this rejection sampling strategy is therefore bound to fail.

Markov Chain Monte Carlo (MCMC) techniques can also be used to draw samples from arbitrary distributions with densities known up to a constant. The density of $F_{X|X_{max},X_{min}}$ is obtained up to a constant multiplier through a simple application of Bayes' theorem. It is proportional to the prior density of F_X multiplied by indicators ensuring that the extrema are respected.

$$\begin{aligned} \mathbb{P}\left(X \mid X_{max}, X_{min}\right) &\propto \mathbb{P}\left(X\right) \mathbb{P}\left(X_{max}, X_{min} \mid X\right) , \\ &\propto \mathbb{P}\left(X\right) \mathbb{I}\left(\max_i \{X_i\} = X_{max}\right) \mathbb{I}\left(\min_i \{X_i\} = X_{min}\right) . \end{aligned} \tag{10}$$

However, once again, this distribution is zero everywhere in $\mathbb{R}^p$, except in a (p-2) dimensional subspace where the min and max constraints are met. A rejection sampler targeting (10) will also fail, and any naive MCMC algorithm will not yield samples from $F_{X|X_{max},X_{min}}$. We therefore approximate the constraint by replacing the likelihood term $\mathbb{P}\left(X_{max}, X_{min} \mid X\right)$ with two narrow independent normal distributions around the minimum and maximum of X. This "softens" the conditional distribution,

$$\mathbb{P}\left(X \mid X_{max}, X_{min}\right) \propto \mathbb{P}\left(X\right) \mathcal{N}\left(X_{max} \mid \max_i \{X_i\}, \epsilon^2\right) \mathcal{N}\left(X_{min} \mid \min_i \{X_i\}, \epsilon^2\right) , \tag{11}$$

where $\mathcal{N}\left(x \mid \mu, \sigma^2\right)$ denotes the density of a normal distribution with mean μ and variance σ^2 evaluated at x. For small ϵ, this seems to be a reasonable approximation enabling the use of MCMC techniques.

This approximation to $F_{X|X_{max},X_{min}}$ remains a difficult distribution to sample from. We illustrate the constraint in a 3-dimensional setting in Figure 5. The MCMC algorithm must travel efficiently along the six edges of the allowed subspace, and navigate corners when the index of the extremum components change. Hamiltonian Monte Carlo (HMC) has shown a a remarkable ability to navigate complicated distributions, including distributions where the typical set has "pinch points" of strong curvature [2], similar to the "corners" in $F_{X|X_{max},X_{min}}$. We therefore used HMC as implemented by the Stan probabilistic programming language [3] to obtain draws from $F_{X|X_{max},X_{min}}$.

HMC's efficient sampling relies on gradient information in order to move towards regions of high probability. The normal likelihood (11) softened the extrema constraints, but the maximum and minimum functions also remove information from the gradient. The partial derivative of the log-likelihood of the maximum term with respect to X_i is proportional to

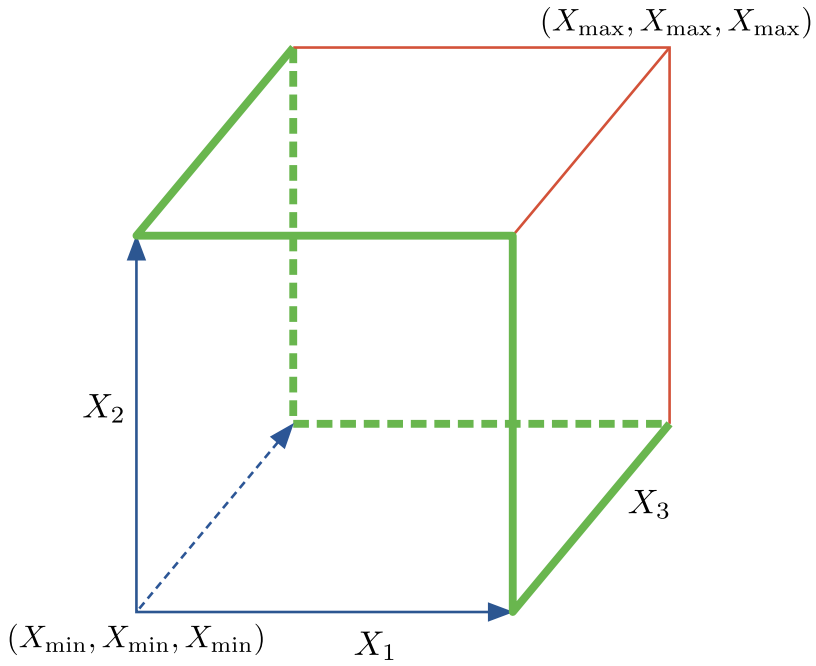

Figure 5: *With three variables X_1, and X_2 and X_3, $F_{X|X_{max},X_{min}}$ resides in the one-dimensional six-sided loop shown with thicker green lines. This is a 1D manifold embedded in 3D space, and possessing sharp corners, making it difficult for most MCMC algorithms to explore.*

$$\frac{\partial \log \mathcal{N}\left(X_{max} \mid \max_i \{X_i\}, \epsilon^2\right)}{\partial X_i} \propto (X_{max} - X_i)\, \mathbb{I}\left\{\arg\max_j (X_j) = i\right\} . \tag{12}$$

In other words, the gradient pulls the maximum of the current sample towards X_{max}, and ignores all other components. This makes it difficult for HMC to efficiently explore scenarios where other components are the maximum.

In order to assist the HMC algorithm, we make another approximation. We replace the max and min functions in (11) with the smoothmax and smoothmin functions, defined on real inputs $x_1, \ldots, x_p$ as:

$$\begin{aligned} \text{smoothmax}\,(x_1, \ldots, x_p; k) &= \frac{1}{k} \log\left(\sum_{i=1}^{p} e^{k x_i}\right) , \\ \text{smoothmin}\,(x_1, \ldots, x_p; k) &= -\,\text{smoothmax}\,(-x_1, \ldots, -x_p; k) . \end{aligned} \tag{13}$$

As the sharpness parameter k goes to infinity, smoothmax becomes the maximum, and smoothmin becomes the minimum. When smoothmax replaces max and smoothmin replaces min, there is a small price in precision due to the approximation, but there is an important benefit: the gradient is now informative for all components of X:

$$\frac{\partial \log \mathcal{N}\left(X_{max} \mid \text{smoothmax}\left(X_{1:p}; k\right), \epsilon^2\right)}{\partial X_i} \propto \left(X_{max} - \text{smoothmax}\left(X_{1:p}; k\right)\right) \frac{e^{kX_i}}{\sum_{j=1}^{p} e^{kX_j}} \,. \tag{14}$$

These modifications make HMC a viable algorithm to efficiently draw samples from the constrained posterior. Setting k and ϵ is a compromise between exactness and efficiency; we found $k = 10$ and $\epsilon = 0.1\,^\circ$ C to perform well for our application.

Henceforth, we will refer to this use of HMC and a smoothmax approximation to the target distribution as SmoothHMC. SmoothHMC provides a generally applicable algorithm to draw from a multivariate distribution conditionally on the observed minimum and maximum of its components.

4.2 Illustration of Hamiltonian Monte Carlo with Smoothmax Approximation

We demonstrate SmoothHMC's ability to obtain draws from $F_{X|X_{max},X_{min}}$ in a simplified setting where the distribution function of $F_{X|X_{max},X_{min}}$ can be derived analytically and also computed easily. In our application, F_X is the posterior predictive multivariate normal distribution $T_{miss} \mid T_{nearby}$ obtained from nearby measurements, with mean and marginal variance evolving smoothly from one prediction to the next. To parallel this, we specify a random vector X with each component X_i normally distributed, and with sinusoidal means and variances, but without any correlations between them, so as to avoid a combinatorial explosion when obtaining the distribution function analytically:

$$\begin{aligned}
X_i &\sim \mathcal{N}\left(\mu_i, \sigma_i\right) \,, \\
X_i &\perp X_j \;\forall\, i \neq j \,, \\
X_{max} &= \max_i \{X_i\} \,, \\
X_{min} &= \min_i \{X_i\} \,, \\
\mu_i &= 10 + \sin\left(2\pi i/50\right) \,, \\
\sigma_i &= 0.1 + \cos^2\left(2\pi i/50\right) \,, \\
i &= 1, 2, \ldots, 100 \,.
\end{aligned} \tag{15}$$

The unconstrained distribution of X_i is illustrated in Figure 6(a). In this example, we aim to sample from the distribution of X_i subject to the observation that $X_{max} = 12.5$ and $X_{min} = 8.8$. We chose this example to have an analytically and computationally constrained distribution $F_{X|X_{max},X_{min}}$ (see derivation in Appendix 9) so that we can verify the correctness of the imputations. The marginal quantiles of the

analytical posterior are shown in Figure 6(b).

To obtain samples from $F_{X|X_{max},X_{min}}$, we use the implementation of Hamiltonian Monte Carlo provided by Stan. In Stan, the user specifies a probabilistic data-generating process for the observed data, based on parameters and latent variables with accompanying priors. Stan then compiles this model into a custom C++ program that implements posterior sampling using HMC. We implement two Stan models to draw from $F_{X|X_{max},X_{min}}$. The Stan model code for both are available from the GitHub account of the first author. The first model implements (11), with the narrow normal likelihood term around the maximum and minimum, while the second model includes the smoothmax and smoothmin approximations (13) to the maximum and minimum functions. For each Stan model, we obtain 4 HMC chains each with 10,000 warm-up samples followed by 10,000 samples. The quantiles of the samples obtained without the smoothmax approximation are shown in Figure 6(c). By default, Stan initizializes each X_i uniformly at random between -2 and 2, and for most variables, the algorithm remains stuck near the initial values. Furthermore, most samples do not conform to the constraints imposed by the observed X_{min} and X_{max} values, which further invalidates these results. However, once we replace the maximum function with the smoothmax function, with quantiles shown in Figure 6(d), Stan is able to draw samples that respect the observed extrema. Furthermore, a visual comparison of the analytical quantiles in Figure 6(a) and the Stan sample quantiles in Figure 6(d) confirms that this sampling algorithm delivers a close approximation of the marginal distribution of each variable X_i in $F_{X|X_{max},X_{min}}$.

We may also visually verify that Stan samples correctly from the *joint* distribution of any combination of variables. We do this for a pair of variables, X_{23} and X_{52}, with results shown in Figure 7. In that figure, the central scatterplot shows the 40,000 Stan samples obtained using the smoothmax approximation. Superimposed on the scatterplot are a contour plot (with dash-dotted lines) of the probability distribution function of the analytical conditional distribution $F_{X|X_{max},X_{min}}$ of X_{23} and X_{52} when neither X_{23} nor X_{52} is one of the extrema, multiplied by the probability of that being the case, which is available analytically. This can be compared to the contour plot (solid lines) of the same probability distribution function obtained through a kernel density estimator of the subset of Stan samples where neither X_{23} and X_{52} is the minimum or maximum, using a normal kernel with bandwidth 0.2, and multiplied by the proportion of samples where that is the case. The kernel estimates are divided by the integrated probability mass of the kernel that is inside of the boundaries, in order to reduce boundary effects. The thin black dotted line are one kernel bandwidth away from the X_{min}/X_{max} boundaries. Outside the thin black dotted line, the kernel density estimates are less trustworthy. The four histograms around the scatter plot are of the Stan samples where one of the variables is an extremum, weighted so as to integrate to the fraction of samples that satisfy that condition. For example, the top histogram is of X_{23} for samples where X_{52} is the maximum, and integrates

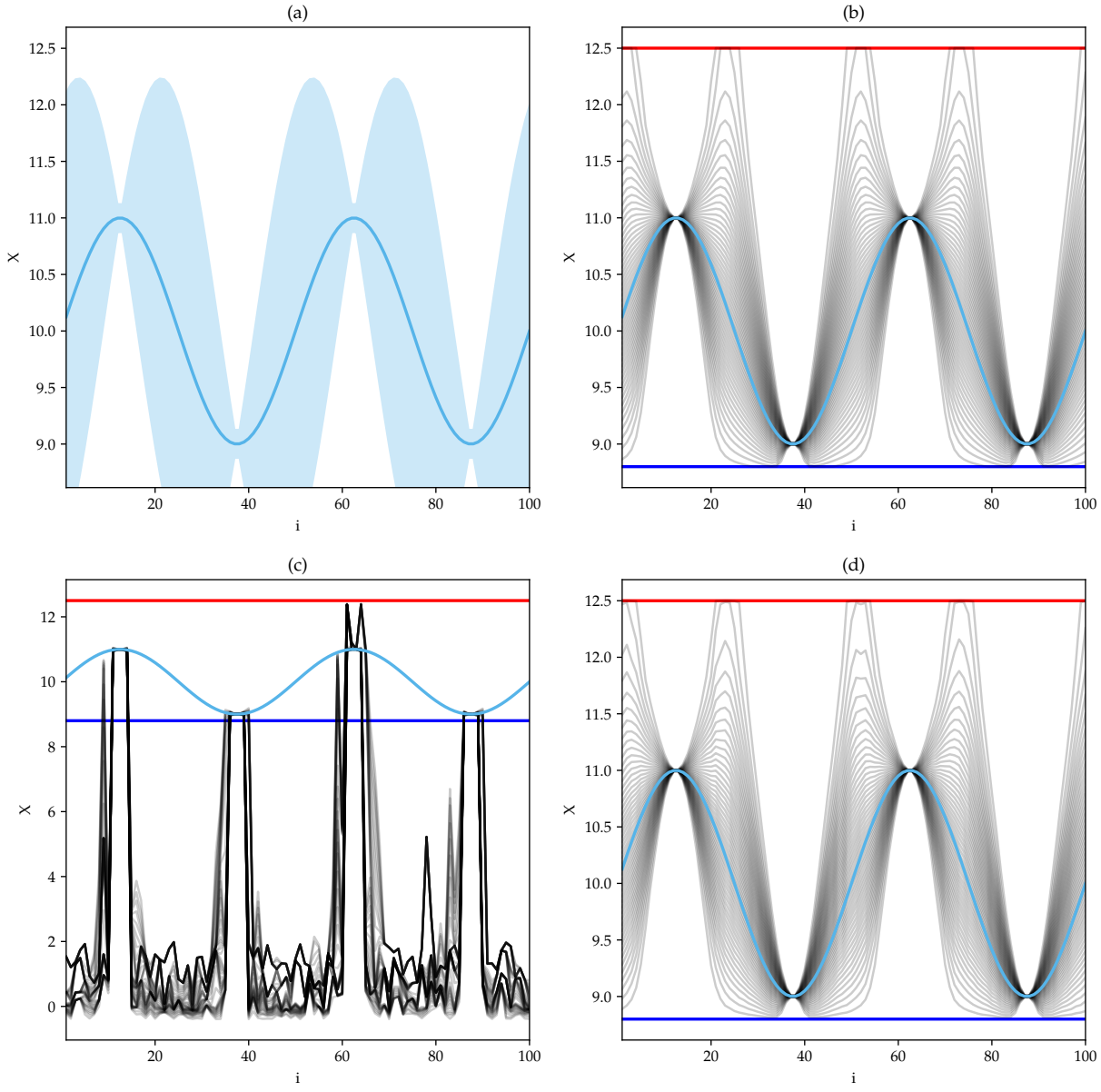

Figure 6: *(a) Prior distribution of* X_i *displayed as mean function with 2 SD envelope; (b) Quantiles of the analytically derived posterior* $F_{X|X_{max},X_{min}}$ *conditioned on* X_{min} *and* X_{max}*, with prior* μ_i *shown in blue; (c) Quantiles of the samples drawn from* $F_{X|X_{max},X_{min}}$ *using Stan without the* smoothmax *approximation, with prior* μ_i *shown in blue; (d) Quantiles of the samples drawn from* $F_{X|X_{max},X_{min}}$ *using Stan with the* smoothmax *approximation, with prior* μ_i *shown in blue.*

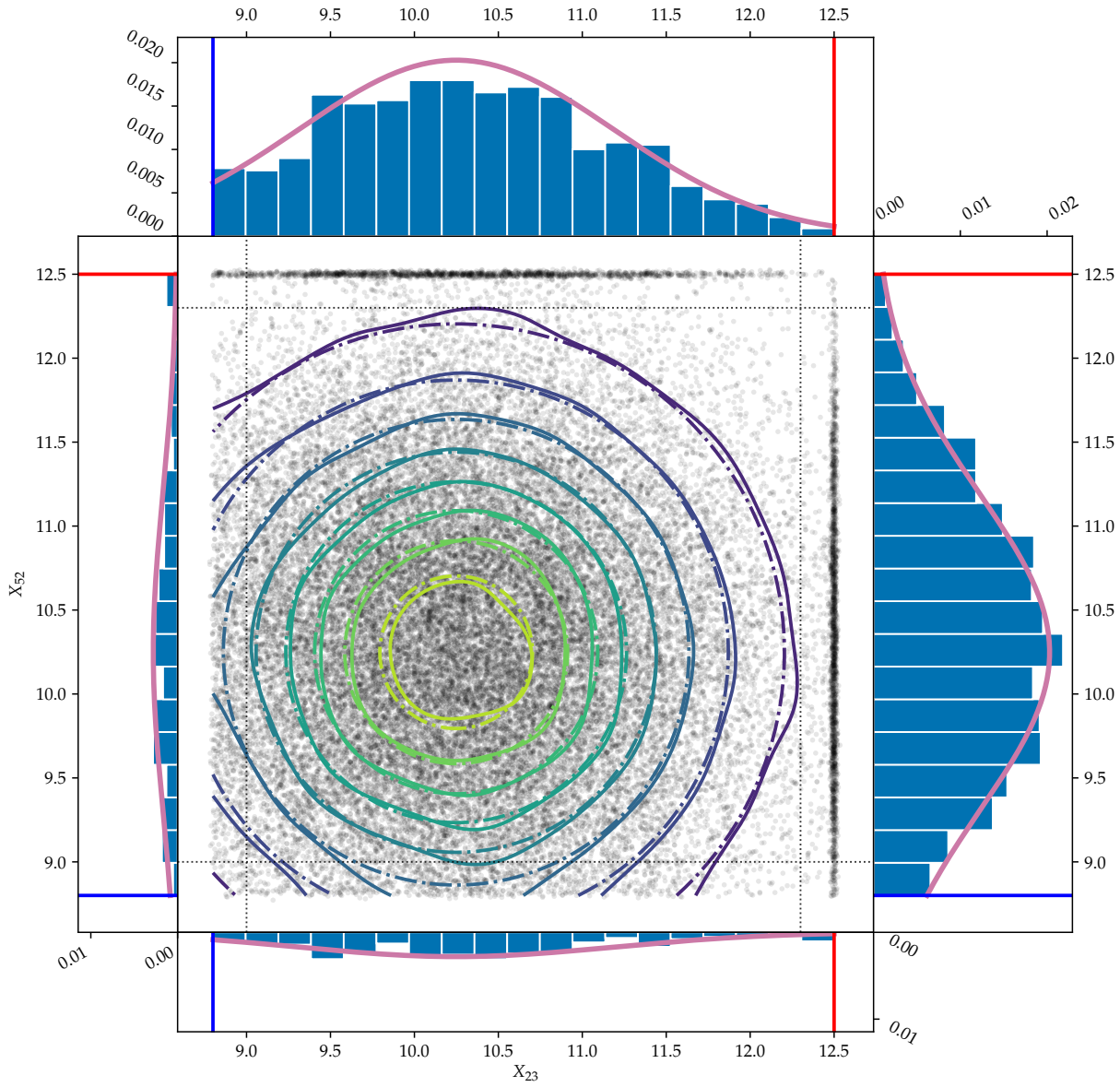

Figure 7: *Comparison of the joint distribution of* X_{23} *and* X_{52} *obtained analytically and from Stan samples.*

to the fraction of samples where that is the case. The super-imposed pink line is that of a truncated normal probability distribution function multiplied by the probability of the satisfied condition. For example, the pink line over the top histogram integrates to $\mathbb{P}_{\bullet 52}$. Lastly, the blue and red lines are X_{min} and X_{max} respectively. There is a close match between the contours of the analytical joint distribution function (dash-dotted lines) and of the kernel density estimate of the Stan samples. Each of the four histogram of samples where X_{23} or X_{52} occupies the minimum or maximum position matches the corresponding analytical distribution function. This visual comparison of the sample and analytical distributions shows that Stan is yielding a good approximation of a sample drawn from the true $F_{X|X_{max},X_{min}}$ in this example. We did not examine the behavior of the sampling algorithm for the joint distribution of more than two variables due to the difficulty of visualizing such a distribution, but we see no reason to suspect that the algorithm suffers from pathological behaviors that do not appear in these univariate and bivariate inspections.

4.3 Smoothmax Temperature Model

Armed with the SmoothHMC algorithm implemented in Stan, we now return to the problem of imputing hourly temperature measurements. To impute the missing temperatures, we need to draw from the posterior distribution $T_{miss} \mid T_{nearby}, T_n, T_x$. Bayes' theorem conditional on T_{nearby} gives

$$\mathbb{P}\left(T_{miss} \mid T_{nearby}, T_n, T_x\right) = \frac{\mathbb{P}\left(T_n, T_x \mid T_{nearby}, T_{miss}\right) \mathbb{P}\left(T_{miss} \mid T_{nearby}\right)}{\mathbb{P}\left(T_n, T_x \mid T_{nearby}\right)}. \tag{16}$$

The second term in the numerator is the posterior obtained in Section 3 now acting as a prior. The denominator is a normalizing constant. The first term in the numerator is either zero or one, indicating whether T_{miss} satisfies the constraint imposed by the observed T_n and T_x. Therefore, the posterior distribution takes a similar form to (10), which motivates the use of SmoothHMC.

A small leap of faith is needed to accept that SmoothHMC's success in a toy example in Section 4.2 will extend to this application. There are three important differences between the toy example and the temperature time series model. Firstly, F_X is now a multivariate normal distribution with strong correlations obtained as the posterior distribution of a Gaussian process in Equation (9). Secondly, instead of a single minimum and maximum, we observe extrema for every 24 hour period. Thirdly, we allow for the mean temperature to be different at different locations, and so the imputed temperatures are shifted by an additional parameter μ_{miss}, to which we attach a vague prior. To summarize, the probabilistic model that we

wish to draw posterior imputations of T_{miss} from is given by:

$$\begin{aligned}
\mu_{miss} &\sim \mathcal{N}\left(0, 10^2\right) && \text{(vague prior on mean temperature)} \\
T_{miss|nearby} \equiv T_{miss} \mid T_{nearby} &\sim \mathcal{N}\left(\mu_{miss|nearby}, \Sigma_{miss|nearby}\right) && \text{(posterior from } T_{nearby} \text{ becomes prior)} \\
T_{miss} &= \mu_{miss} + T_{miss|nearby} \\
(T_x)_d &= \max\left\{T_{miss,i}, \text{ for all } i \text{ such that } t_{miss,i} \in (t_{d-1}^{meas}, t_d^{meas}]\right\}, \\
(T_n)_d &= \min\left\{T_{miss,i}, \text{ for all } i \text{ such that } t_{miss,i} \in (t_{d-1}^{meas}, t_d^{meas}]\right\}.
\end{aligned} \tag{17}$$

To sample from this model, we modify it with the smoothmax approximation to the maximum, and a normal likelihood:

$$\begin{aligned}
(T_x)_d &\sim \mathcal{N}\left(\underset{i \in (t_{d-1}^{meas}, t_d^{meas}]}{\text{smoothmax}}\{T_{miss,i}; k = 10\}, 0.1^2\right), \\
(T_n)_d &\sim \mathcal{N}\left(\underset{i \in (t_{d-1}^{meas}, t_d^{meas}]}{\text{smoothmin}}\{T_{miss,i}; k = 10\}, 0.1^2\right).
\end{aligned} \tag{18}$$

A few samples from this imputation procedure are shown in Figure 4(c). From May 28, 2015 to June 1, 2015, hourly temperatures are imputed at Waterloo Airport, using the hourly temperature measurements from nearby stations to inform the course of the temperatures, and using the daily minima and maxima extracted from the hourly measurements to constrain the imputed temperatures. One can verify visually that the imputations respect the T_n and T_x constraints, reaching but not exceeding each extreme on each day. Since we actually have hourly data for Waterloo, yet fed our algorithm only the daily extremes, we can also plot the hidden temperatures (in black), and see how faithfully the imputations reproduce them. We see that the imputations indeed track the true measurements closely. On May 31st, we can see that the imputations capture two possibilities: the T_x record *could* have been set early in the 24-hr period, or at its end. This success demonstrates that SmoothHMC is capable of imputing temperature time series from the constrained posterior distribution $T_{miss} \mid T_{nearby}, T_n, T_x$.

5 Model diagnostics

5.1 Variogram

We can visually inspect our model by plotting temporal and spatial semi-variograms. The semi-variogram of a stationary spatio-temporal function $Y(\mathbf{x}, t)$ is a function of the spatial lag $\mathbf{h}$ and the temporal lag r[see

for example 13, chapter 6]:

$$\gamma(\mathbf{h}, r) = \frac{1}{2}\mathbb{E}\left[(Y(\mathbf{x}, t) - Y(\mathbf{x}+\mathbf{h}, t+r))^2\right] = \text{var}(Y(\mathbf{x}, t)) - \text{cov}((Y(\mathbf{x}, t)), Y(\mathbf{x}+\mathbf{h}, t+r)) \,. \tag{19}$$

For a Gaussian Process model, with a stationary kernel $k(\mathbf{h}, r) = k(\mathbf{x}, \mathbf{x}+\mathbf{h}, t, t+r)$ this can be expressed in terms of the observation noise σ_ϵ and kernel function $k(\cdot, \cdot)$, as

$$\gamma(\mathbf{h}, r) = \sigma_\epsilon^2 + k(0, 0) - k(\mathbf{h}, r) \,. \tag{20}$$

From the data, the semi-variogram can also be estimated empirically, be averaging the square differences of any two observations that are separated by $\mathbf{h}$ in space, and r in time (or, in practice, within half a bin width of $\mathbf{h}$ and r). By comparing the empirical variogram to the variogram of our fitted $\mathcal{GP}$ model, we obtain a visual diagnosis of the model.

In our Iowa example, there are only four possible locations. For each location, we plot the empirical temporal variogram $\widehat{\gamma}(0, r)$. For any pair of stations separated by $\mathbf{h}$ (fixed), we can also plot $\widehat{\gamma}(\mathbf{h}, r)$. We then overlay the model's semi-variogram obtained through equation (20), resulting in Figure 8.

We notice that the variogram of the simple `SExSE` model tracks the empirical variogram well at short lags, but fails to capture the diurnal cycle, and the fit degrades at long lag. We attempt to improve the model in Section 6.

5.2 Error and expected error

The variogram gives us a visual diagnostic of the overall model fit. To quantify the model's predictive ability in the Iowa example, we compare the posterior mean temperature to the withheld truth, and obtain the empirical mean squared error as

$$\text{MSE}\left(\text{err} \mid T_{\text{nearby}}, T_x, T_n\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}\left(T_{\text{miss},i} \mid T_{\text{nearby}}, T_x, T_n\right) - T_{\text{miss},i}\right]^2 \,. \tag{21}$$

Equation (21) is for the final predictions obtained using nearby hourly temperatures and local daily maxima and minima. A similar diagnostic can be computed for the intermediary predictions, which exclude the local T_x and T_n information. At that stage, we are not concerned with any overall bias in the predicted temperatures, so we instead compute the sample variance of the errors as

$$\text{var}\left(\text{err} \mid T_{\text{nearby}}\right) = \text{var}_i\left\{\mathbb{E}\left(T_{\text{miss},i} \mid T_{\text{nearby}}\right) - T_{\text{miss},i}\right\} \,. \tag{22}$$

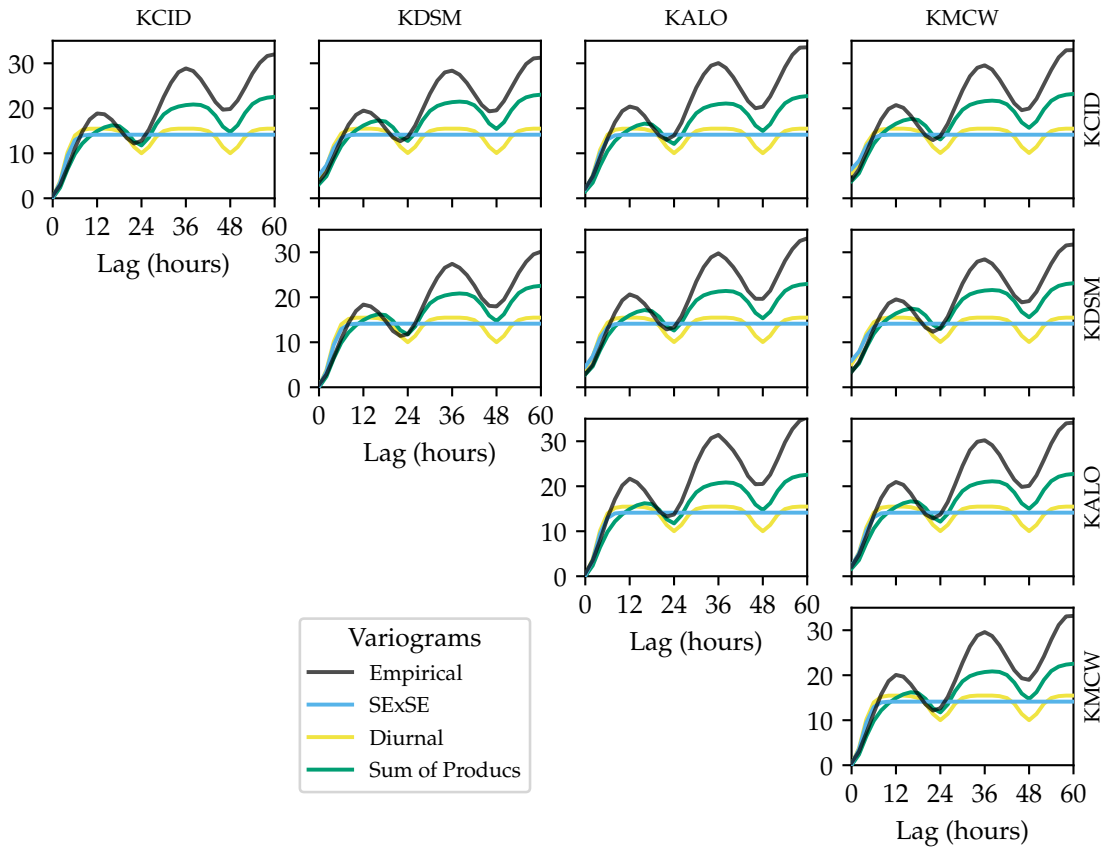

Figure 8: *Semi-variogram*

Table 1: *Model diagnostics for three Gaussian process covariance functions.*

Model	Log Likelihood	Var(err)	$\mathbb{E}$(Var(err))	MSE(err)	$\mathbb{E}$(MSE(err))
$k_{\texttt{SExSE}}$	-55,614	1.59	0.88	1.12	0.44
$k_{\texttt{SESE}_{24}}$	-54,472	1.63	0.97	1.12	0.69
$k_{\texttt{sumprod}}$	-45,944	1.32	1.19	1.04	0.81

For our purposes, it isn't sufficient for the spatio-temporal model to yield good predictions; we also require a good estimate of its own accuracy. We estimate the expected MSE and predictive variance by sampling K random draws T^k_{miss} from the posterior distribution, again conditioned firstly on just T_{nearby} after fitting the spatio-temporal Gaussian process model, and then additionally on T_{nearby}, T_x and T_n after incorporating the local data using Stan. The draws are obtained from the posterior multivariate normal distribution in the first case, and the MCMC samples obtained through Stan in the second case. We then evaluate the variance or MSE between the samples and the posterior mean as

$$\begin{aligned}
\mathbb{E}\left(\text{var}\left(\text{err} \mid T_{\text{nearby}}\right)\right) &\approx \frac{1}{K}\sum_{k=1}^{K} \text{var}_i\left\{T^{(k)}_{\text{miss},i} - \mathbb{E}\left(T_{\text{miss},i} \mid T_{\text{nearby}}\right)\right\} \\
\mathbb{E}\left(\text{MSE}\left(\text{err} \mid T_{\text{nearby}}, T_x, T_n\right)\right) &\approx \frac{1}{K}\sum_{k=1}^{K} \text{MSE}_i\left\{T^{(k)}_{\text{miss},i} - \mathbb{E}\left(T_{\text{miss},i} \mid T_{\text{nearby}}, T_x, T_n\right)\right\}
\end{aligned} \tag{23}$$

When evaluating models, we want the errors to be small, and so the empirical error variance and MSE to be low. A well-calibrated model should also have the expected error variances $\mathbb{E}\left(\text{var}\left(\text{err} \mid \cdot\right)\right)$ close to their empirical values.

These diagnostics for our first spatio-temporal model, the product of squared exponentials, are found in the first row of Table 1. The empirical error variance using only nearby measurements is already fairly low, with typical errors of order $\sqrt{1.59} = 1.26\,^\circ$ C. Incorporating the local measurements reduces it further to $\sqrt{1.12} = 1.06\,^\circ$ C. However, the model is overly optimistic, and the expected errors underestimate the true errors.

6 Improving the basic model

In this section, we develop more sophisticated Gaussian process models than the simple product of squared exponential kernels. We then assess whether these models improve the variogram and the predictive diag-

nostic measures that we developed in the previous sections.

The most salient feature of the empirical variogram that is not captured by the SExSE model is the oscillation with a 24-hour period. It is intuitively clear that the diurnal cycle induces this periodic covariance, and that our model should be improved by incorporating this feature. Gaussian process models allow for periodic components of the covariance, for example the periodic squared exponential kernel, which we will use with a 24-hour period

$$k_{24}(t, t') = \sigma_{24}^2 \exp\left[-\frac{2}{\ell_{24}^2}\sin^2\left(\pi\frac{t-t'}{24\text{ hrs}}\right)\right]. \tag{24}$$

We modify the spatiotemporal model by adding this diurnal component to it, with its own spatial decay kernel $k_{space24}$ (with the same specification as k_{space} in (1), and again with variance parameter fixed to 1):

$$k_{SESE_{24}}(\mathbf{x}, \mathbf{x}', t, t') = k_{time}(t, t') \cdot k_{space}(\mathbf{x}, \mathbf{x}') + k_{24}(t, t') \cdot k_{space24}(\mathbf{x}, \mathbf{x}') + k_{\mu}(\mathbf{x}, \mathbf{x}'). \tag{25}$$

We also develop a more complex model, which breaks up k_{time} into short-term, medium-term and long-term correlation components, each with their own spatial decay.

$$\begin{aligned} k_{sumprod}(\mathbf{x}, \mathbf{x}', t, t') = {} & k_{time1}(t, t') \cdot k_{space1}(\mathbf{x}, \mathbf{x}') && \text{(short-term variation)} \\ & + k_{time2}(t, t') \cdot k_{space2}(\mathbf{x}, \mathbf{x}') && \text{(medium-term variation)} \\ & + k_{time3}(t, t') \cdot k_{space3}(\mathbf{x}, \mathbf{x}') && \text{(long-term variation)} \\ & + k_{24}(t, t') \cdot k_{space24}(\mathbf{x}, \mathbf{x}') && \text{(diurnal cycle)} \\ & + k_{\mu}(\mathbf{x}, \mathbf{x}') && \text{(station mean)} \end{aligned} \tag{26}$$

Each of k_{time1}, k_{time2}, and k_{time3}, is a rational quadratic kernel

$$k_{RQ}(t, t') = \sigma^2\left(1 + \frac{(t-t')^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{27}$$

which is accompanied by its spatial decay kernel, specified as a squared exponential covariance with variance fixed at 1. This more complicated kernel therefore has $3 \times 3 \times 2 + 2 \times 2 = 22$ free parameters, in addition to the noise parameter σ_ϵ^2.

We now have three competing Gaussian process models, with covariance functions k_{SExSE}, $k_{SESE_{24}}$, and $k_{sumprod}$ respectively. We can compare them in three ways. Firstly, the marginal log-likelihood is the quantity maximized by the parameter fitting procedure in (8). The maximized log-likelihood can be found in the second column of Table 1, and we see that the more complex models indeed yield a much higher log-

likelihood, promising a better model fit which should yield better predictions. Secondly, we compare the variance of the error in the predicted temperatures specified in (22) when withholding all the data from a test station. Averaged over all of 2015, this is given in the third column, and shows more mixed results. The diurnal model $k_{\mathtt{SESE}_{24}}$ performs slightly worse than the simple $k_{\mathtt{SExSE}}$ model, and $k_{\mathtt{sumprod}}$ only yields a small improvement. Thirdly, we can reintroduce the daily minima and maxima from the withheld station, and compare the mean squared error specified in (21) for predictions at the test station. Results in the fifth column show even more modest improvements for the more complex models. However, the more complex model give better estimates of their own inaccuracy: the expected errors reported by the model are closer to the empirical errors for predicting the witheld time series.

We interpret these results as a reminder that prediction accuracy using Gaussian process is sensitive to model specification when extrapolating, but fairly insensitive to the model when interpolating [14]. Our imputations interpolate the temperatures from nearby stations, further aided by the constraints imposed by the daily T_n and T_x measurements, which could explain why the choice of model does not seem to have a large impact on the performance of our imputation procedure. This insensitivity can be seen as reassuring, as it shows robustness against model misspecification.

7 Imputed summary statistics

Figure 4(d) shows the imputations produced under the $k_{\mathtt{sumprod}}$ kernel (26). This is the primary output of our imputation method, and the results are promising. Firstly, just like in the toy example presented in 4.2, the individual imputations meet the three constraints imposed by the measured minimum and maximum. Each day, the imputations stay between T_n and T_x, and the temperatures always drop to T_n and rise to T_x at some time of the day. The imputations reflect the uncertainty in the time at which the extrema are reached. Notably, on some days, the posterior distribution of the warmest (or coldest) time is bimodal. For example this can be seen on May 31st, where some imputations reach T_x at the start of the measurement window, while others reach it at the end. We view as a particular strength of our approach that the imputations are able to capture this bimodality.

These imputations however are not the final aim of our analysis. Rather, our stated goal is to undo, or at least account for, the sensitivity of summary statistics to measurement time, for example the average T_x in Figure 2. Equipped with these imputations, is it possible to infer what the value of the summary statistic would have been for different measurement hours? This seems to be possible as demonstrated in Figure 9, which shows the same summary statistic as in Figure 2 applied to the imputations as well as the (witheld) hourly data at Waterloo Airport. It can be seen that the imputed summary statistics track within about

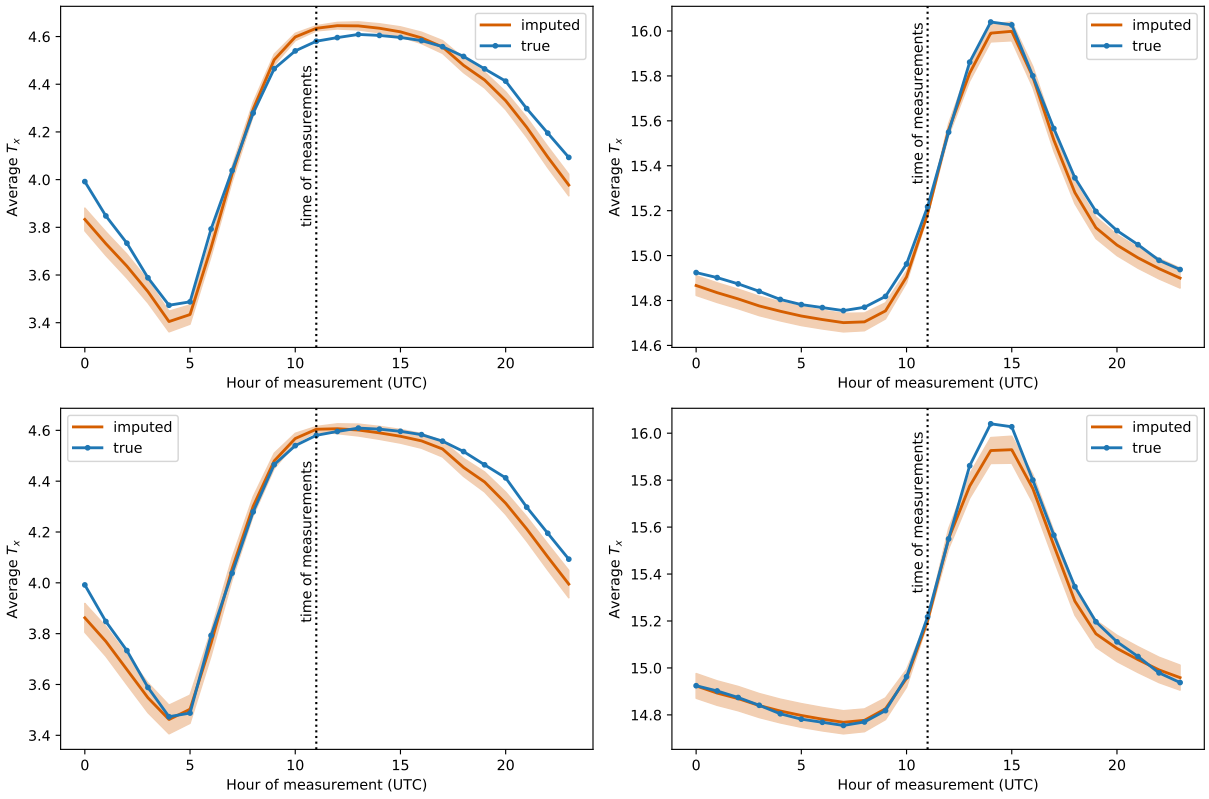

Figure 9: *Average maximum daily temperature summary statistic obtained under varying hour of measurement from withheld Waterloo Airport data (shown in blue), and from imputations of the withheld data (shown in orange with 2 SD envelope).*

0.1 ° C of the true values, while the posterior standard deviation gives a fair estimate of the imputation error.

8 Inference on measurement hour

Our analysis thus far has focused on the case where the hour of measurement `hour` is known in advance. This is an unrealistic assumption in practice, and so inference on `hour` is a desirable feature. It is conceptually straightforward to modify the measurement model (18) with a uniform prior on `hour`. However, because we obtain our imputations in ten-day windows, in most windows precise information about `hour` will not be available, as moving the measurement time one hour earlier or later rarely affects the measured T_n and T_x. Furthermore, `hour` affects which observations are attributed to each day's measurements. This effect is discontinuous (observations suddenly jump from one day to the next) and non-differentiable, and so Hamiltonian Monte Carlo becomes unviable. This issue is similar to that caused by the non-differentiability of the minimum and maximum functions. We therefore do not consider the introduction of a uniform prior on `hour` in Stan to be feasible.

Our procedure allows us to obtain imputation samples of T_{miss} conditional on $T_{\text{nearby}}, T_n, T_x$ and `hour`. If we do so for $\texttt{hour} = 1, 2, \ldots, 24$, is there information available in these samples to infer `hour`? We will examine sample imputations to answer this question. Figure 10 shows mean imputation for temperatures over nine days starting on February 27, 2015. The orange line is the mean using only nearby temperatures (shifted by a constant to match the true temperatures), while the green line is additionally conditional on T_n and T_x; the true temperatures are shown in grey. The top plot shows the imputation under the correct daily measurement time (11:00 UTC-6), while the bottom plot is under an incorrect measurement time (23:00 UTC-6). The first unsurprising observation is that assuming an incorrect measurement time can lead to wildly inaccurate imputations. But we then also notice that assuming the wrong time also causes the mean constrained imputation to depart further from the unconstrained imputation (that is, the green and orange lines are further apart). This can be interpreted as an indication of an incompatibility between T_{nearby} and the daily extremes, caused by assuming the wrong `hour`. To quantify this discrepancy, we propose to calculate the probability of the mean constrained imputation under the unconstrained posterior given by (9):

$$\begin{aligned} \mu(\texttt{hour}) &\equiv \mathbb{E}\left(T_{\text{miss}} \mid T_{\text{nearby}}, T_n, T_x, \texttt{hour}\right) \text{ (the mean imputed temperature)}, \\ \delta_{\texttt{hour}} &\equiv \mathbb{P}\left(T_{\text{miss}} = \mu(\texttt{hour}) \mid T_{\text{nearby}}\right). \end{aligned} \tag{28}$$

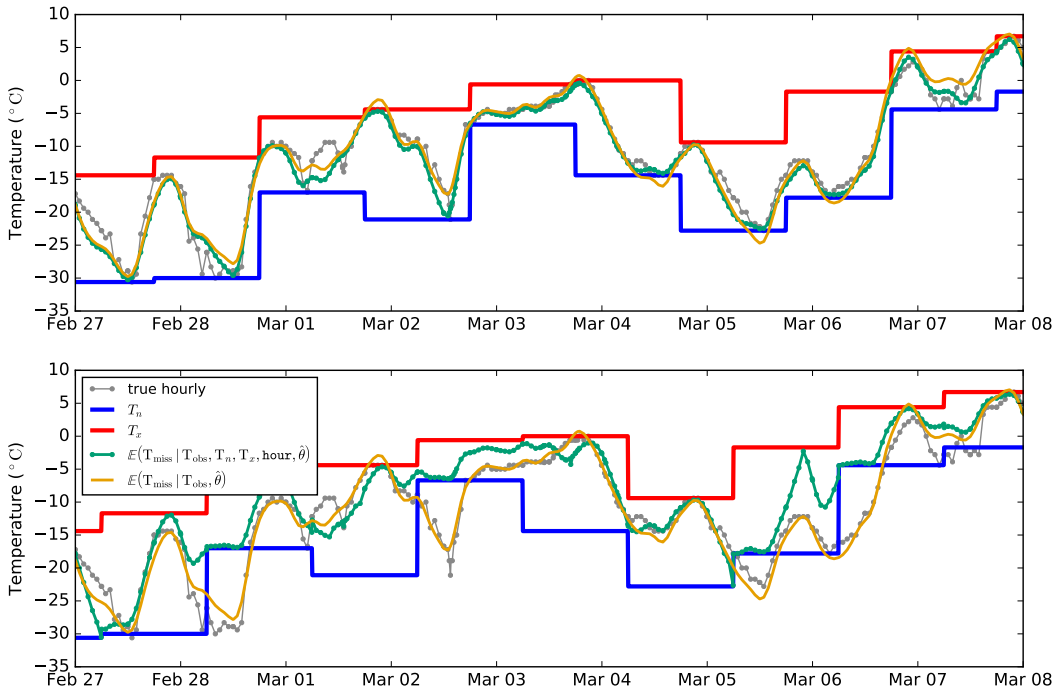

Figure 10: *A sample window showing constrained and unconstrained imputations assuming (top) the correct measurement hour (11:00 UTC-6), and (bottom) the wrong measurement hour (23:00 UTC-6). Assuming the wrong measurement time drives the constrained mean imputation away from the unconstrained mean imputation.*

Our intuition is that $\delta_{\texttt{hour}}$ will drop sharply when the wrong `hour` is assumed, and we may be able to infer the true `hour` by maximizing $\delta_{\texttt{hour}}$.

9 Conclusion

Climatological research relies on the ability to track small changes over long periods. For this reason, the bias induced by the measurement time that we demonstrate in Section 1.1 could lead to wrong estimates and conclusions regarding long-term trends in temperature records. We reformulated the source of this bias as a missing data problem, and imputed the missing hourly temperatures at the weather station using posterior samples from a spatiotemporal Gaussian process model. The model allows the combination of information from the measured daily minimum ($\mathrm{T_n}$) and maximum ($\mathrm{T_x}$) temperatures, and from measurements of hourly temperatures at nearby meteorological stations. While ours is not a physical model, it is very flexible, and it performs well for the task of interpolating temperatures between nearby locations and

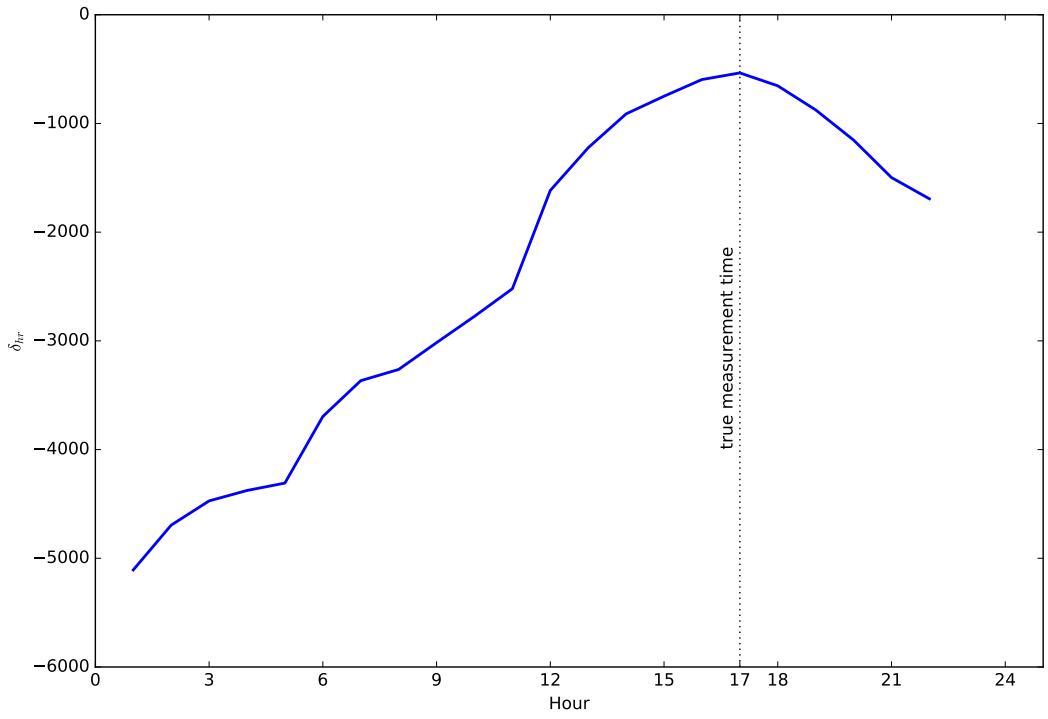

Figure 11: *Discrepancy measure for imputations of temperatures at Waterloo Municipal Airport assuming measurement hours* $\texttt{hour} = 1, 2, \ldots, 24$*. The true hour of measurement is 17, and obtains the highest* $\delta_{\texttt{hour}}$*.*

times. Indeed, more complex covariance functions (with a diurnal component and a sum of short-range and long-range components) showed only modest improvements in the mean squared error of the imputations compared to a withheld hourly temperature record.

Our model accounts for miscalibration and bias in the hourly temperature measurements by assigning a mean parameter to each location, which is given a weak independent prior with no spatial correlation. Therefore, our model only makes predictions at new locations up to a constant shift, and it only extracts information about the trajectory of the temperature time series from each weather station. However, our strategy rests on the assumption that the trajectory is not affected by biases and miscalibration. This assumption is violated for example if the presence of an airport has a very different effect on measured temperatures during the day and during the night, which would introduce bias in the imputations. Our model could be improved in the future with a more complete characterization of how daily temperatures differ systematically between locations.

In order to condition the imputations on the daily T_x and T_n, we developed SmoothHMC, a general algorithm based on Hamiltonian Monte Carlo with a smoothed approximation of the target distribution that can sample from a multivariate distribution conditionally on its observed minimum and maximum. It showed an excellent ability to sample from the conditional distribution in an example where the distribution function can also be obtained analytically. SmoothHMC is the main technical contribution of this paper, and we believe the method could find applications beyond the present setting.

We used this method to obtained imputations of the temperature time series that satisfied the constraints imposed by the measured T_n and T_x. The imputation of withheld temperatures at Waterloo Municipal Airport track the true temperatures, within a root mean square error of $1.02\,^\circ$ C. We view as particularly encouraging that the imputations successfully capture bimodalities of the possible time of the maximum or minimum temperature on days where this time is difficult to infer from the available information.

Future improvements to the imputation strategy would include the inclusion of rounding effects in the measurement model, explicit treatment of non-stationarity due to coastlines or other geographical features, and of altitude differences. Gaussian process modeling allows for much flexibility in the choice of covariance kernels, and improved modeling should lead to more accurate imputations.

The imputed time series are the primary output of this work, but they are intended as a starting point for further analyses motivated by different scientific goals. In particular, summary statistics can be applied to the imputations, such as the average T_x, under different choices of daily measurement hours. Using imputations obtained for the withheld time series at Waterloo airport, we have demonstrated a good ability to recover this information, though the resulting posterior variance seems to underestimate the error. The average T_x is an example of a possible follow-up analysis, chosen mostly as an illustrative proof of concept.

We plan to use this method to compare the average temperature to the average of the measured T_n and T_x for a given location and year, with the former obtained from imputed time series.

Appendix: Derivation of the analytic posterior for toy example

We first derive and compute $F_{X|X_{max},X_{min}}$ for this example. We denote by $f_i(\cdot)$ and $F_i(\cdot)$ the prior probability distribution function and cumulative distribution function of X_i, i.e. the normal PDF and CDF with means and variances given by (15). Let $\mathbb{P}_{ij}$ be the probability that X_i is the minimum of X, and X_j is its maximum. We also define $\mathbb{P}_{i\bullet} = \sum_{j=1}^{100} \mathbb{P}_{ij}$, the probability that X_i is the minimum, and $\mathbb{P}_{\bullet j} = \sum_{i=1}^{100} \mathbb{P}_{ij}$, the probability that X_j is the maximum. The cumulative distribution function of X_i is then given by

$$\mathbb{P}(X_i \leqslant x \mid X_{max}, X_{min}) = \begin{cases} 0 & \text{when } x < X_{min}\,, \\ \mathbb{P}_{i\bullet} + (1 - \mathbb{P}_{i\bullet} - \mathbb{P}_{\bullet i}) \left[\frac{F_i(x) - F_i(X_{min})}{F_i(X_{max}) - F_i(X_{min})}\right] & \text{when } X_{min} \leqslant x < X_{max}\,, \\ 1 & \text{when } x \geqslant X_{max}\,. \end{cases} \tag{29}$$

Meanwhile, $\mathbb{P}_{ij}$ is proportional to

$$f_i(X_{min}) f_j(X_{max}) \prod_{k \neq i,j}^{100} (F_k(X_{max}) - F_k(X_{min}))\,, \tag{30}$$

which we compute for all i, j and renormalize to obtain the 100×100 matrix of probabilities. We sum over its rows and columns to obtain $\mathbb{P}_{\bullet j}$ and $\mathbb{P}_{i\bullet}$. While this algorithm has cubic complexity in the dimensionality p of X, for $p = 100$, it only take seconds to compute the entries of $\mathbb{P}$ and evaluate $\mathbb{P}(X_i \leqslant x \mid X_{max}, X_{min})$ over a range of x. Figure 6(b) shows the analytical quantiles of $F_{X|X_{max},X_{min}}$. Roughly speaking, we see that the prior distribution F_X is stretched to fit between X_{min} and X_{max}.

References

[1] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

[2] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[3] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017.

[4] PM Della-Marta and Heinz Wanner. A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate*, 19(17):4179–4197, 2006.

[5] Jean-François Ducré-Robitaille, Lucie A Vincent, and Gilles Boulet. Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, 23(9):1087–1101, 2003.

[6] David R Easterling, Thomas C Peterson, and Thomas R Karl. On the development and use of homogenized climate datasets. *Journal of climate*, 9(6):1429–1434, 1996.

[7] Thomas R Karl, Claude N Williams Jr, Pamela J Young, and Wayne M Wendland. A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States. *Journal of Climate and Applied Meteorology*, 25(2):145–160, 1986.

[8] Matthew J Menne, Imke Durre, Russell S Vose, Byron E Gleason, and Tamara G Houston. An overview of the Global Historical Climatology Network-Daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897–910, 2012.

[9] Matthew J Menne and Claude N Williams Jr. Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7):1700–1717, 2009.

[10] Matthew J Menne, Claude N Williams Jr, and Russell S Vose. The US Historical Climatology Network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, 90(7):993–1007, 2009.

[11] Thomas C Peterson, David R Easterling, Thomas R Karl, Pavel Groisman, Neville Nicholls, Neil Plummer, Simon Torok, Ingeborg Auer, Reinhard Boehm, Donald Gullett, et al. Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology*, 18(13):1493–1517, 1998.

[12] Joaquin Quinonero-Candela, Carl Edward Rasmussen, and Christopher KI Williams. Approximation methods for gaussian process regression. *Large-scale kernel machines*, pages 203–224, 2007.

[13] Michael Sherman. *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons, 2011.

[14] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[15] Blair Trewin. A daily homogenized temperature data set for Australia. *International Journal of Climatology*, 33(6):1510–1529, 2013.

[16] Lucie A Vincent, Xiaolan L Wang, Ewa J Milewska, Hui Wan, Feng Yang, and Val Swail. A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. *Journal of Geophysical Research: Atmospheres*, 117(D18), 2012.