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

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 and easy to implement Markov Chain Monte Carlo 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 recover the hidden temperatures well. 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 (Menne et al., 2012), 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 (1) changes in the time of observation, (2) changes in the thermometer technology, (3) station relocation, and (4) changes in land use around a station (Menne et al., 2009). 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., Karl et al., 1986; Easterling et al., 1996; Peterson et al., 1998; Ducré-Robitaille et al., 2003; Menne and Williams Jr, 2009; Vincent et al., 2012), 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 appli-

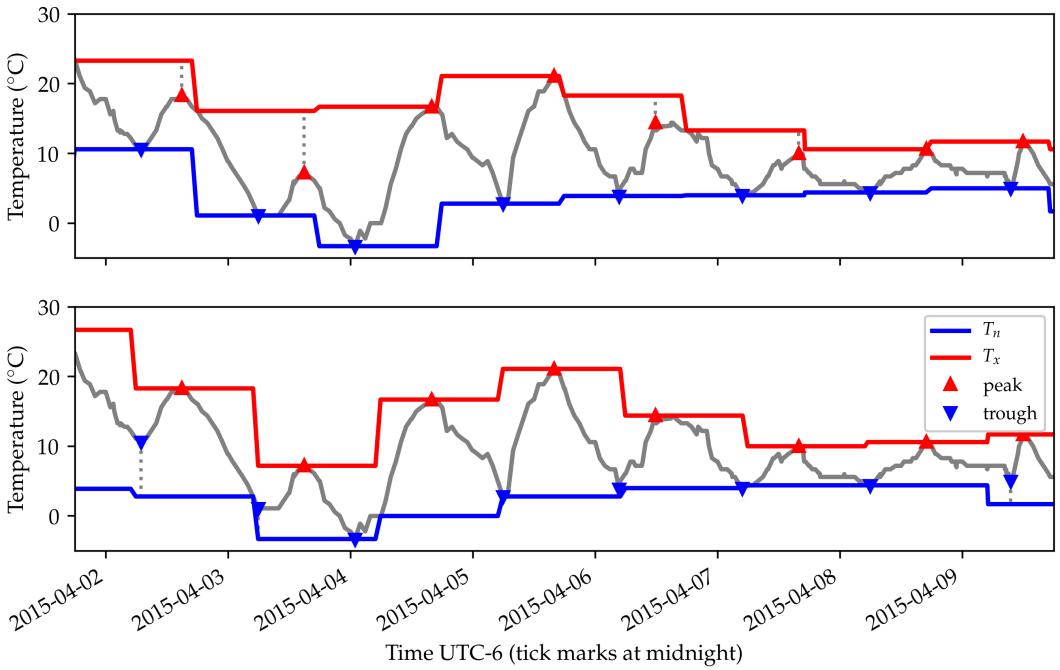

Figure 1: An extract of the temperature measurements from KALO. The blue and red triangles respectively indicate the coldest and warmest temperature of each diurnal cycle. The blue and red lines respectively show the observed maximum and temperature recorded each day at 17:00 (top) or 5:00 (bottom) for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

cations of these methods, however, focus on adjusting the mean state of the data rather than the shape of the distribution (see Della-Marta and Wanner, 2006, 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 (Della-Marta and Wanner, 2006; Trewin, 2013).

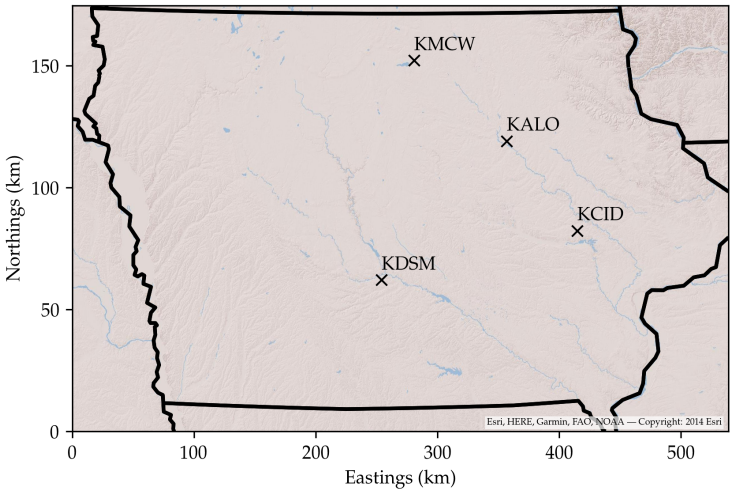

Figure 2: Map of the four airport weather stations in Iowa providing hourly temperature records. Each airport is identified by its ICAO code.

1.1 The Problem

Many historical measurements of daily temperatures are provided as daily maximum and minimum temperatures (T_x and T_n respectively), which ideally measure the peak and trough of each diurnal temperature cycle. T_x and T_n are often recorded by an observer who every 24 hours visits a weather station equipped with a maximum-minimum thermometer, and notes the maximum and minimum registered by the instrument in the last 24 hours. In this section we explain how this measurement practice can cause the T_x and T_n measurements to fail to capture the peaks and troughs of some diurnal cycles. This has long been recognized in the scientific literature; see for example Baker (1975) and references therein.

Figure 1 illustrates the problem with ten days of hourly temperature measurements from the Waterloo Municipal Airport (KALO) weather station in Iowa. Records of these measurements are publicly available in the Integrated Surface Database (ISD), which is a compilation of global weather data containing approximately 14,000 active stations and maintained in the United States by NOAA's National Centers for Environmental Information (NCEI). Figure 2 gives a map of the four Iowa weather stations used as examples throughout this paper. We emulate daily T_x/T_n measurements by dividing the data into 24 hour measurement

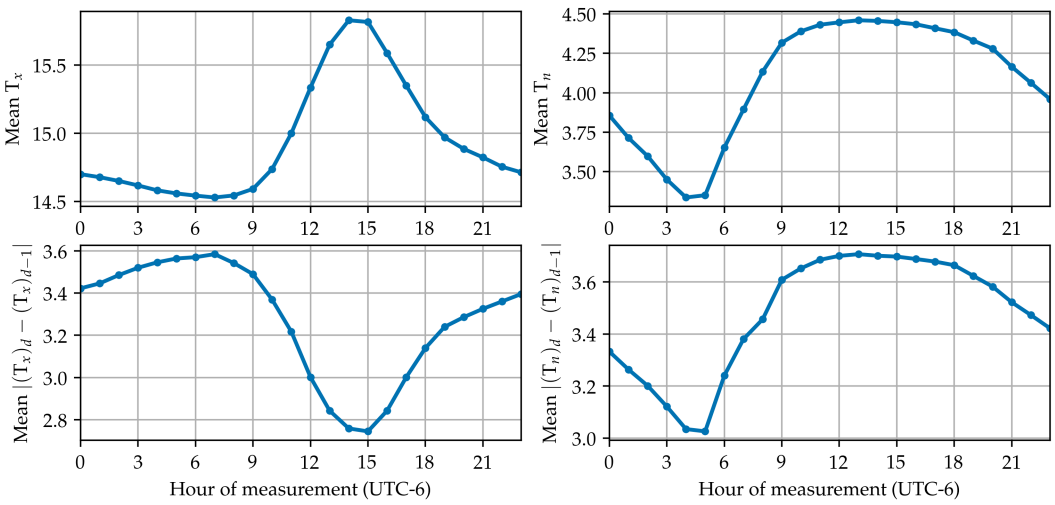

Figure 3: Mean daily T_x (top left) and T_n (top right), and mean absolute daily change in T_x (bottom left) and T_n (bottom right), extracted from hourly temperature records at KALO in 2015, under varying measurement hours of T_x and T_n.

windows, and reporting the minimum and maximum temperature that was recorded in this window. On most days, the measurements successfully capture the peak and trough of the diurnal cycle. But there are also several discrepancies (indicated with vertical dotted lines), 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. A blatant example occurs on April 3rd, where the peak of the diurnal cycle is 7.2°C and occurs at 15:00 (all times are in the UTC-6 time zone, and tick marks are at midnight at the start of each day), but with measurements made at 17:00, the day's T_x record of 16.1°C is reached immediately after the previous day's measurement: a 8.9°C overestimate. 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 would minimize the possibility of the previous or next diurnal cycle setting the measured T_x or T_n. For convenience, however, most observers instead record data at a single daytime hour. 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 statistic 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, as seen in Figure 3. Conversely, the average T_n is colder by over 1°C if T_n is measured at 5:00 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 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 (Menne et al., 2009). Such spurious trends also compromise the study of weather variability, through summary statistics such as the average absolute change in daily temperature maxima and minima from one day to the next, as seen in Figure 3.

1.2 Our Approach

One of our goals is to be able to infer the "true" T_x and T_n peaks and troughs of the diurnal cycle throughout the data records, so as to correct both the variance biases and the spurious trends. This stands in constrast to previous work, which has focused directly on addressing spurious trends. We approach the problem as a missing data problem: if we had access to the full temperature time series at the station rather than just T_x and T_n measured at an arbitrary time, we would be able to retrospectively choose the hour of measurements, to avoid the issues described in Section 1.1. Our idea therefore is to impute the hourly time series of temperatures at the location of the T_x/T_n measurements. In turn, the imputed time series can be used to create accurate summaries of temperature extremes.

Our imputation strategy is to borrow information from the nearby weather stations, usually located at airports, that record the current temperature about once an hour. For this purpose, we use the publicly available temperature data provided by ISD. Although it should be noted that the sampling times are not always equally spaced, we refer to these records as

"hourly" throughout this paper. They often cannot be used directly for climatology, as the weather stations that provide them are not always as carefully documented, calibrated, and situated as the research stations included in the GHCND. For instance, weather stations at locations experiencing a lot of human activity, like airports, may record higher temperatures on average. However, even if mis-calibrated or systematically biased, the time series 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 pooling 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 SmoothHMC, a novel and easy to implement Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution satisfying the above constraints. Our constrained imputations are implemented in the Stan programming language (Carpenter et al., 2017); our code is publicly available on the first author's GitHub account. Compared to a custom implementation, the Stan model code is short and Stan's MCMC samplers are well-optimized, which makes our imputation strategy efficient and easy to reproduce.

2 A First Spatiotemporal Model

In order to pool the information from temperatures measured 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 also decay in time. We model the simultaneous temperatures throughout a region as a Gaussian process, with covariance between two locations $\boldsymbol{x}$ and $\boldsymbol{x}'$

given by the squared exponential (SE) covariance with characteristic lengthscale ℓ_{space} and variance σ^2_{space}:

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

Similarly, the time series of temperatures at a single location can be modeled as a Gaussian process with characteristic timescale ℓ_{time} and variance σ^2_{time}:

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

We combine the spatial and temporal model by multiplying the covariances functions:

$$k(\boldsymbol{x}, \boldsymbol{x}', t, t') = k_{\text{time}}(t, t') \cdot k_{\text{space}}(\boldsymbol{x}, \boldsymbol{x}'). \tag{3}$$

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 $\sigma^2_{\text{space}} = 1$. To allow for systematic differences between stations, we add a mean temperature parameter $\mu_{\text{station}[i]}$ 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:

$$\begin{aligned} \mathrm{T}_{ij} &= \mu_j + f(\boldsymbol{x}_j, t_{ij}) + \epsilon_{ij}\,, \quad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}\big(0, \sigma^2_\epsilon\big) \\ f(\boldsymbol{x}_j, t_{ij}) &\sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}', t, t'))\,. \end{aligned} \tag{4}$$

The noise term captures measurement error and micro-fluctuations occuring on time scales much shorter than ℓ_{time}. At stations j that only provide daily T_x and T_n records, we denote the time of the d^{th} daily measurement by t_d^{meas}, and approximate the T_x and T_n observation

respectively as the maximum or minimum temperatures at a discretized set of times t_{ij} inside of $(t_{d-1}^{\text{meas}}, t_d^{\text{meas}}]$:

$$\begin{aligned}(\mathrm{T}_x)_{dj} &= \max\left\{\mathrm{T}_{ij}, \text{ for all } i \text{ such that } t_{ij} \in (t_{d-1}^{\text{meas}}, t_d^{\text{meas}}]\right\}, \\ (\mathrm{T}_n)_{dj} &= \min\left\{\mathrm{T}_{ij}, \text{ for all } i \text{ such that } t_{ij} \in (t_{d-1}^{\text{meas}}, t_d^{\text{meas}}]\right\},\end{aligned} \tag{5}$$

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 Quinonero-Candela et al. (2007) review sparse approximations to Gaussian processes from a machine learning perspective, while Banerjee et al. (2008) 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 $\sigma_\mu = 10\,^\circ\text{C}$, which can be incorporated into the Gaussian process with an additional term

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

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

$$\mathrm{k}_{\texttt{SExSE}}(\boldsymbol{x}, \boldsymbol{x}', t, t') = k_{\text{time}}(t, t') \cdot k_{\text{space}}(\boldsymbol{x}, \boldsymbol{x}') + k_\mu(\boldsymbol{x}, \boldsymbol{x}'), \tag{7}$$

which we denote $\mathrm{k}_{\texttt{SExSE}}$ to distinguish it from the covariance functions developed later in Section 6. Our model thus has four free parameters, σ_{time}, ℓ_{time}, ℓ_{space} and σ_ϵ, which we

	Covariance Function		
Parameter	$\mathtt{k_{SExSE}}$	$\mathtt{k_{SESE_{24}}}$	$\mathtt{k_{sumprod}}$
σ_ϵ (°C)	0.4	0.4	0.2
σ_{time} (°C)	3.7	3.1	$0.5, 0.9, 4.4$
ℓ_{time} (hr)	2.7	2.8	$0.3, 1.9, 8.9$
ℓ_{space} (km)	176	154	$10, 59, 370$
α_{time}			$0.3, 1.1, 0.3$
σ_{24} (°C)		2.4	2.7
ℓ_{24} (hr)		0.7	0.8
ℓ_{space24} (km)		1414	785

Table 1: Fitted parameters for each specification of the Gaussian process covariance function. For $\mathtt{k_{sumprod}}$ (27), the parameters of the short-term, medium-term, and long-term components are separated by commas. Notice how shorter timescales ℓ_{time} are associated with shorter lengthscales ℓ_{space} by the fitted covariance function.

fit by maximizing the marginal likelihood of T, the complete 2015 temperature time series provided at the four Iowa weather stations:

$$\widehat{\sigma}_{\text{time}}, \widehat{\ell}_{\text{time}}, \widehat{\ell}_{\text{space}}, \widehat{\sigma}_\epsilon = \underset{\sigma_{\text{time}}, \ell_{\text{time}}, \ell_{\text{space}}, \sigma_\epsilon}{\arg\max} \left\{ \mathbb{P}(\mathrm{T} \mid \sigma_{\text{time}}, \ell_{\text{time}}, \ell_{\text{space}}, \sigma_\epsilon) \right\}. \tag{8}$$

The fitted covariance values are found in Table 1.

3 Predictions Using Nearby Data

After optimizing the parameters of the spatio-temporal covariance (7), we use the model (4)—fitted to the data from nearby stations with full time series—to provide time series predictions at the station that only collects T_x and T_n data. Gaussian processes give closed-form expressions for the posterior distribution of the predicted temperatures. We denote the temperatures we wish to impute as T_{miss} at times t_{miss} and location $\boldsymbol{x}_{\text{miss}}$ and those observed at nearby stations as $\mathrm{T}_{\text{nearby}}$, at times t_{nearby} and locations X_{nearby}. Under the spatio-temporal model (4), T_{miss} and $\mathrm{T}_{\text{nearby}}$ are jointly multivariate normal, with mean zero and covariance given by $\mathtt{k_{SExSE}}(\boldsymbol{x}, \boldsymbol{x}', t, t')$. Standard results for conditioning within multivariate normals

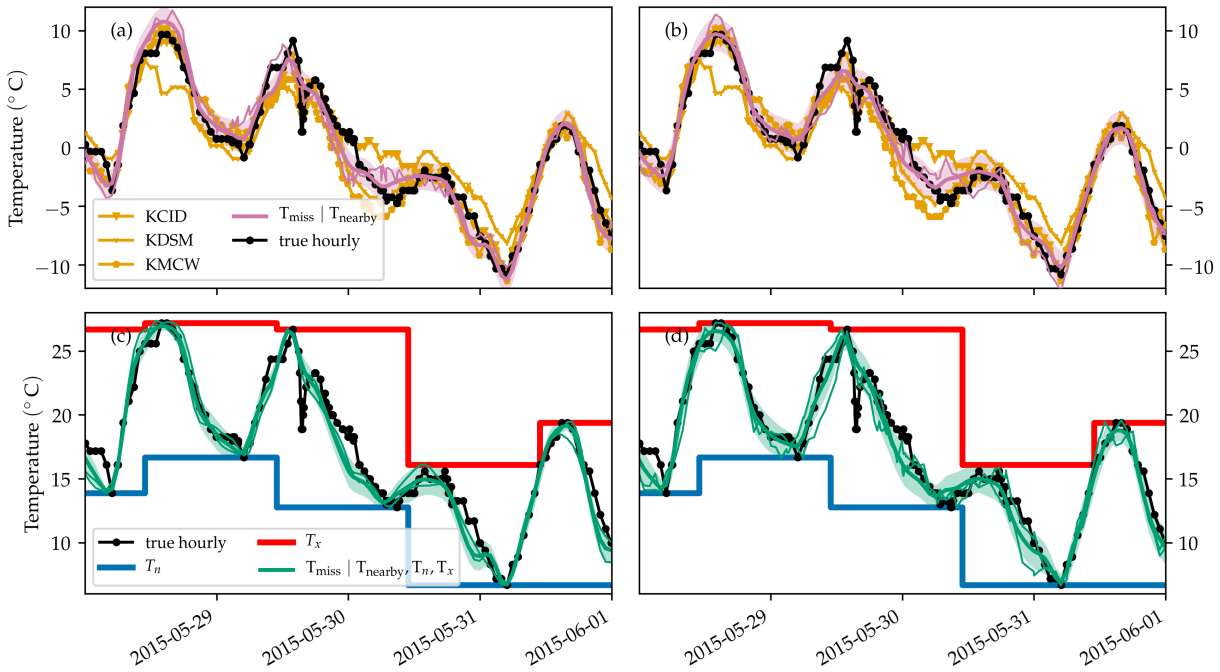

Figure 4: Imputations of the temperature time series at Waterloo Municipal Airport (KALO) between May 28, 2015 and June 1, 2015 (a) using only nearby data and the product of squared exponentials model; (b) using only nearby data and the sum of products model; (c) incorporating T_n and T_x measurements under the product of squared exponentials model; and (d) incorporating T_n and 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.

then yield:

$$\begin{aligned}
&\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}} \sim \mathcal{N}\left(\mu_{\text{miss}|\text{nearby}}, \Sigma_{\text{miss}|\text{nearby}}\right), \text{ with} \\
&\quad \mu_{\text{miss}|\text{nearby}} = \mathbb{E}(\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}}) = \mathbf{K}_{\text{miss,nearby}} \boldsymbol{\Sigma}^{-1}_{\text{nearby,nearby}} \mathrm{T}_{\text{nearby}}, \text{ and} \\
&\quad \Sigma_{\text{miss}|\text{nearby}} = \mathrm{var}(\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}}) = \boldsymbol{\Sigma}_{\text{miss,miss}} - \mathbf{K}_{\text{miss,nearby}} \boldsymbol{\Sigma}^{-1}_{\text{nearby,nearby}} \mathbf{K}^{\intercal}_{\text{miss,nearby}}.
\end{aligned} \tag{9}$$

All covariance matrices can be derived from the model. For example, the ij^{th} entry of $\mathbf{K}_{\text{miss,nearby}} = \mathrm{cov}(\mathrm{T}_{\text{miss}}, \mathrm{T}_{\text{nearby}})$ is given by $\mathrm{k}_{\texttt{SExSE}}(\boldsymbol{x}_{\text{miss}}, X_{\text{nearby}}[j], t_{\text{miss}}[i], t_{\text{nearby}}[j])$, where $X_{\text{nearby}}[j]$ gives the location of the jth observation, and $t_{\text{nearby}}[j]$ its time. The two $\boldsymbol{\Sigma}$ matrices have an additional σ^2_ϵ diagonal component for measurement noise.

In Figure 4(a), we show an example of predictions obtained from this spatio-temporal

model. We withheld temperature measurements from KALO (shown in black), and then used data from the three remaining stations (KCID, KDSM and KMCW, shown in orange) to predict the 2015 temperature time series At KALO. To speed up computations, we process 73 days of data at a time, with 48 days overlapping between adjacent prediction windows so that predictions can always be made away from the edge of the prediction window (except at the start and end of the year). The predictions can be seen to combine information from the three other stations, giving less weight to KDSM, which is further away from KALO. We will discuss the quality of these predictions in more detail in Section 5, after completing the exposition of our imputation strategy.

4 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 $X_{\max} = \max_i\{X_i\}$ and minimum $X_{\min} = \min_i\{X_i\}$ 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:

$$\mathbb{P}(X \mid X_{\max}, X_{\min}) \propto \mathbb{P}(X)\, \mathbb{I}\Big\{\max_i\{X_i\} = X_{\max}\Big\}\, \mathbb{I}\Big\{\min_i\{X_i\} = X_{\min}\Big\}. \tag{10}$$

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. Consequently, out-of-the-box generic MCMC algorithms targeting (10) will not successfully converge to $F_{X|X_{\max},X_{\min}}$. We therefore loosen the constraint by replacing the likelihood term $\mathbb{P}(X_{\max}, X_{\min} \mid X)$ with two narrow independent normal distributions around the minimum and maximum of X:

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

where $\mathcal{N}(x \mid \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2 evaluated at x. For small ϵ, this is 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 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 remarkable ability to navigate complicated distributions, including distributions where the typical set has "pinch points" of strong curvature (Betancourt, 2017), similar to the "corners" in $F_{X|X_{\max},X_{\min}}$. 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

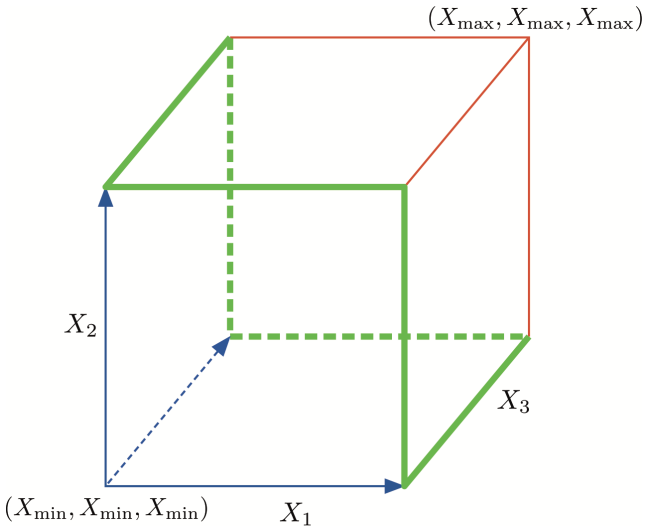

Figure 5: With three variables X_1, and X_2 and X_3, $F_{X|X_{\max},X_{\min}}$ resides in the one-dimensional six-sided loop shown with thicker green lines. This is a 1D manifold embedded in 3D space, and possessing sharp corners, making it difficult for most MCMC algorithms to explore.

log-likelihood of the maximum term with respect to X_i is proportional to:

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

The gradient pulls the maximum of the current MCMC state towards $X_{\max}$, and ignores all other components. This makes it difficult for HMC to efficiently explore scenarios where other components set 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 approaches the maximum, and smoothmin approaches the minimum. This substitution costs a small price in accuracy due to the approximation, but there is an important benefit: the gradient is now informative for

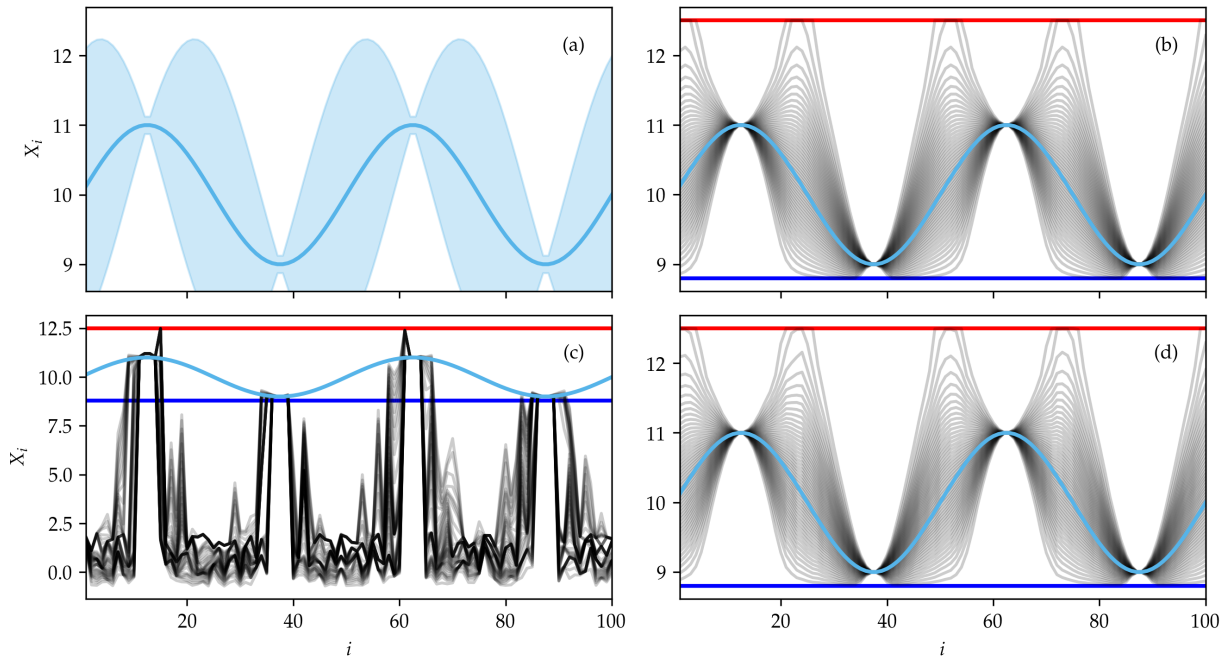

Figure 6: (a) Prior distribution of X_i displayed as mean μ_i (shown in every subplot to ease comparison) with 2 SD envelope; (b) Quantiles of the analytically derived posterior $F_{X|X_{\max},X_{\min}}$ conditioned on $X_{\min}$ (dark blue line) and $X_{\max}$ (red line); (c) Quantiles of the samples drawn from $F_{X|X_{\max},X_{\min}}$ using HMC (without the smoothmax approximation); (d) Quantiles of the samples drawn from $F_{X|X_{\max},X_{\min}}$ using SmoothHMC.

all components of X:

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

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 $\epsilon = 0.1\,^{\circ}\mathrm{C}$ to perform well for our application.

Henceforth, we 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.

4.1 Demonstration of SmoothHMC

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 computed easily. In our application, F_X is the posterior predictive multivariate normal distribution $\mathrm{T}_{\mathrm{miss}} \mid \mathrm{T}_{\mathrm{nearby}}$ 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 $F_{X|X_{\max},X_{\min}}$ analytically:

$$\begin{aligned} &X_i \overset{\perp}{\sim} \mathcal{N}(\mu_i, \sigma_i)\,, \quad i = 1, \ldots, 100\,, \\ &\mu_i = 10 + \sin(2\pi i/50)\,, \quad \sigma_i = 0.1 + \cos^2(2\pi i/50)\,. \\ &X_{\max} = \max_i\{X_i\} \quad \text{and} \quad X_{\min} = \min_i\{X_i\}\,. \end{aligned} \tag{15}$$

The unconstrained distribution of X_i is shown 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 $X_{\min} = 8.8$. An analytical derivation of the marginals of $F_{X|X_{\max},X_{\min}}$ is provided in Appendix A, and its quantiles shown in Figure 6(b).

To obtain samples from $F_{X|X_{\max},X_{\min}}$, we use the implementation of HMC provided by the Stan probabilistic programming language (Carpenter et al., 2017). 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 efficiently implements posterior sampling using HMC. We implement two Stan models to draw from $F_{X|X_{\max},X_{\min}}$; code for both is available from the GitHub account of the first author. The first model implements (11), with a narrow normal likelihood term around the maximum and minimum, while the second model also uses the smoothmax approximation (13). 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. Most samples do not conform to the constraints imposed by the observed $X_{\min}$ and $X_{\max}$ values, which invalidates these imputations. However, once we replace the maximum function with the smoothmax function, with quantiles shown in Figure 6(d), SmoothHMC is able to draw samples that respect the observed extrema. Furthermore, a visual comparison of the analytical quantiles in Figure 6(a) and the SmoothHMC 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 also visually verify that SmoothHMC 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. There is a close match between the contours of the analytical joint distribution function (dash-dotted contour lines) and of the kernel density estimate (solid contour lines) of the SmoothHMC 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 well. This visual comparison of the sample and analytical distributions shows that SmoothHMC 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.

4.2 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 $\mathrm{T}_{\mathrm{miss}} \mid \mathrm{T}_{\mathrm{nearby}}, \mathrm{T}_n, \mathrm{T}_x$. Bayes' theorem

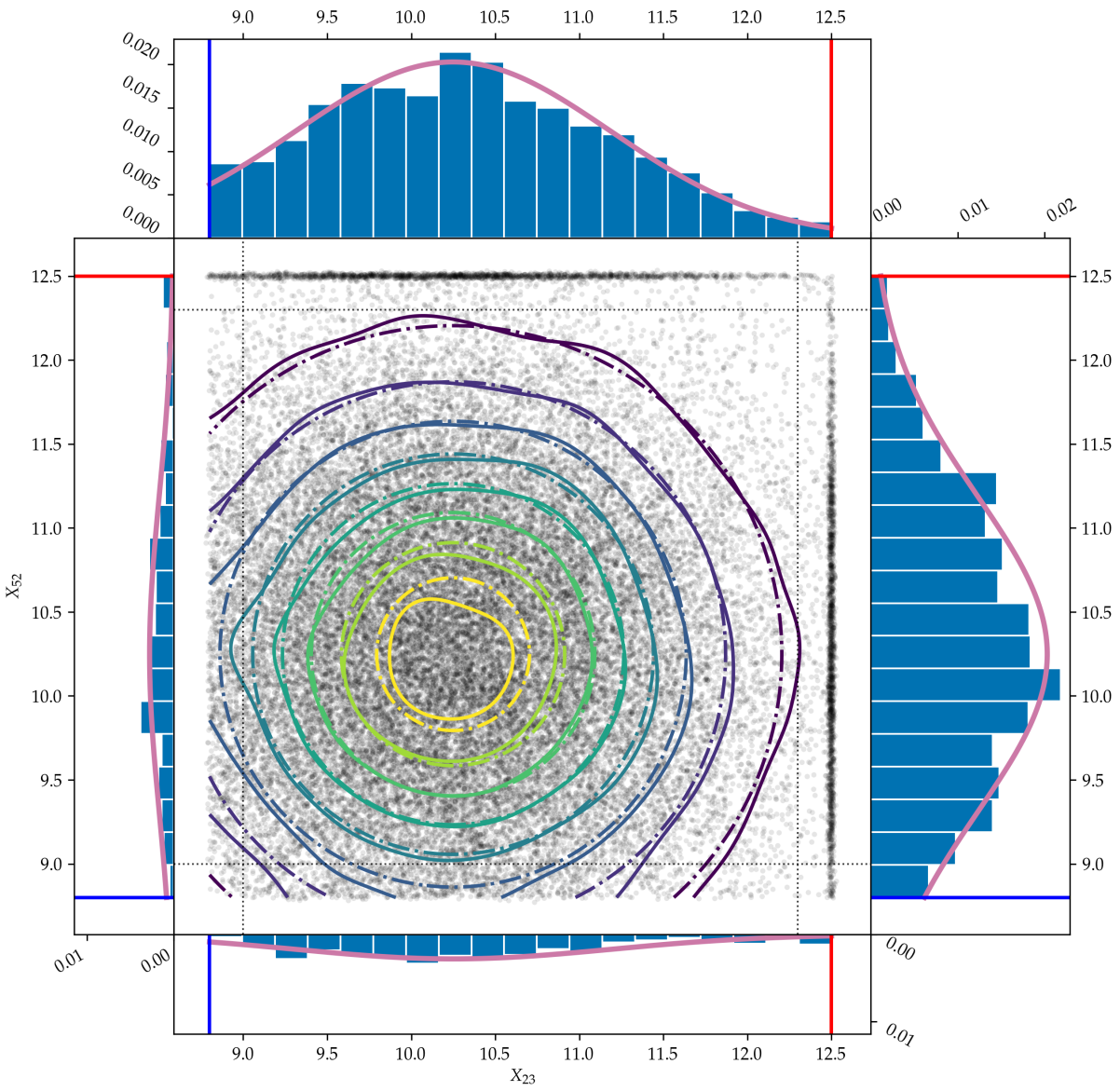

Figure 7: Comparison of the joint joint PDF of X_{23} and X_{52} obtained analytically and from SmoothHMC samples. The central scatterplot shows the 40,000 SmoothHMC samples. Superimposed thereon are a contour plot (dash-dotted) of the joint marginal PDF of $F_{X|X_{\max},X_{\min}}$ for X_{23} and X_{52}, and a contour plot (solid lines) of kernel density estimates for the subset of SmoothHMC samples where neither X_{23} or X_{52} is the min or max, obtained with a normal kernel with bandwidth 0.2 (estimates are divided by the integrated mass of the kernel that is inside of the $X_{\min}/X_{\max}$ boundaries). The dotted lines are one bandwidth away from the $X_{\min}/X_{\max}$ boundaries, beyond which kernel density estimates are less reliable. The four histograms around the scatter plot are of the SmoothHMC samples adjacent to their x-axis, when one of the variables is an extremum. For example, the top histogram is of X_{23} for samples where X_{52} is the max, while the super-imposed pink line is the (truncated normal) marginal PDF of X_{23} if it is neither the max nor the min, times the probability that X_{52} is the max. Blue and red lines indicate $X_{\min}$ and $X_{\max}$ respectively.

conditional on $\mathrm{T}_{\text{nearby}}$ gives

$$\mathbb{P}(\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}}, \mathrm{T}_n, \mathrm{T}_x) = \frac{\mathbb{P}(\mathrm{T}_n, \mathrm{T}_x \mid \mathrm{T}_{\text{nearby}}, \mathrm{T}_{\text{miss}})\, \mathbb{P}(\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}})}{\mathbb{P}(\mathrm{T}_n, \mathrm{T}_x \mid \mathrm{T}_{\text{nearby}})}\,. \tag{16}$$

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 4.1 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 (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:

$$\begin{aligned}
\mathrm{T}_{\text{miss}} &= \mu_{\text{miss}} + \mathrm{T}_{\text{miss|nearby}} \quad \text{with} \quad \mu_{\text{miss}} \sim \mathcal{N}\big(0, 10^2\big)\,, \text{ and} \\
\mathrm{T}_{\text{miss|nearby}} &= \mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}} \sim \mathcal{N}\big(\mu_{\text{miss|nearby}}, \Sigma_{\text{miss|nearby}}\big) \\
(\mathrm{T}_x)_d &= \max\big\{\mathrm{T}_{\text{miss},\,i}, \text{ for all } i \text{ such that } t_{\text{miss}\,,i} \in (t_{d-1}^{\text{meas}},\, t_d^{\text{meas}}]\big\}\,, \\
(\mathrm{T}_n)_d &= \min\big\{\mathrm{T}_{\text{miss},\,i}, \text{ for all } i \text{ such that } t_{\text{miss}\,,i} \in (t_{d-1}^{\text{meas}},\, t_d^{\text{meas}}]\big\}\,.
\end{aligned} \tag{17}$$

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

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

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 KALO, using the hourly temperature measurements from nearby stations to inform the course of the temperatures, and constraining the imputations within the T_x and T_n extracted from the withheld time series at 11:00 each day. Imputations are obtained in nine day windows for computational reasons, with three days of overlap between adjacent windows so each imputation can be made at least three days away from the window's edges. 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 KALO, 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. This success demonstrates that SmoothHMC is capable of imputing temperature time series from the constrained posterior distribution $\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}}, \mathrm{T}_n, \mathrm{T}_x$.

5 Model Diagnostics

5.1 Variogram

Model fit can be visually inspected by plotting temporal and spatial semi-variograms. The semi-variogram of a stationary spatio-temporal function $Y(\boldsymbol{x}, t)$ is a function of the spatial lag $\boldsymbol{h}$ and the temporal lag r (see for example Sherman, 2011, chapter 6):

$$\gamma(\boldsymbol{h}, r) = \frac{1}{2}\,\mathbb{E}\big[(Y(\boldsymbol{x}, t) - Y(\boldsymbol{x} + \boldsymbol{h}, t + r))^2\big] \tag{19}$$

For a Gaussian Process model, with a stationary covariance function $k(\boldsymbol{h}, r) = k(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{h}, t, t + r)$, this can be expressed as:

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

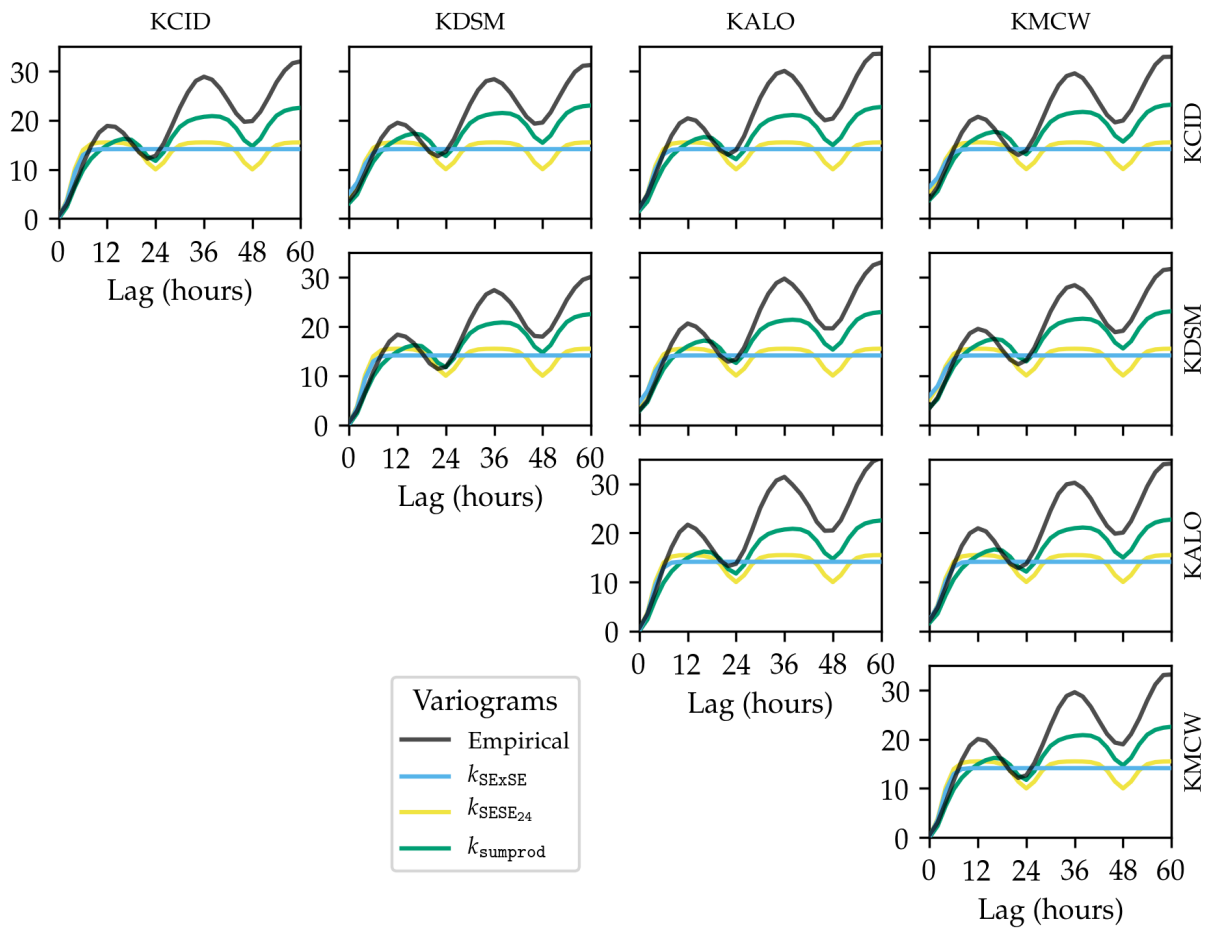

Figure 8: Semi-variograms of the temperature temperature time series at four Iowa weather stations, each labeled by its ICAO code. The empirical semi-variograms are shown in black, and the fitted variograms for the three covariance models proposed in this paper are shown in color. The temporal semi-variograms are shown on the diagonal, while the off-diagonal plots show the semi-variograms as a function of time lag for a fixed distance $\boldsymbol{h}$ equal to the distance between the two stations.

From the data, the variogram can also be estimated empirically, by averaging the square differences of any two observations that are separated by $\boldsymbol{h}$ in space, and r in time (in practice, time lags are binned). By comparing the empirical variogram to the variogram of the fitted covariance, 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 $\widehat{\gamma}(0, r)$. Then, for each pair of stations separated by $\boldsymbol{h}$ (fixed), we can also plot the estimate $\widehat{\gamma}(\boldsymbol{h}, r)$. We overlay the model's semi-variogram from equation

Model	log-likelihood	var(err) (22)	$\widehat{\text{var}}$(err) (23)	MSE (21)	$\widehat{\text{MSE}}$ (24)
$\texttt{k}_{\texttt{SExSE}}$	-55,614	1.59	0.88	1.12	0.44
$\texttt{k}_{\texttt{SESE}_{24}}$	-54,472	1.63	0.97	1.12	0.69
$\texttt{k}_{\texttt{sumprod}}$	-45,944	1.32	1.19	1.04	0.81

Table 2: Model diagnostics for three Gaussian process covariance functions.

(20), resulting in Figure 8. For each variogram, we have removed the effect of the k_μ covariance, which would shift the variogram between two stations by a large arbitrary constant. Correspondingly, we subtract the mean of each observed time series before obtaining the empirical variogram.

We notice that the variogram of the model with product covariance (7) tracks the empirical variogram well at short lags, but fails to capture the periodicity in the empirical variogram, and the fit degrades at long lag. We 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 (MSE) for N predictions as:

$$\frac{1}{N}\sum_{i=1}^{N}[\mathrm{T}_{\mathrm{miss},i} - \mathbb{E}(\mathrm{T}_{\mathrm{miss},i} \mid \mathrm{T}_{\mathrm{nearby}}, \mathrm{T}_x, \mathrm{T}_n)]^2 . \tag{21}$$

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

$$\mathrm{var}(\mathrm{err}) = \mathrm{var}_i\{\mathrm{T}_{\mathrm{miss},i} - \mathbb{E}(\mathrm{T}_{\mathrm{miss},i} \mid \mathrm{T}_{\mathrm{nearby}})\} . \tag{22}$$

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 error variance expected by the model by sampling random draws $\mathrm{T}_{\mathrm{miss}}^{(k)}$, $k = 1, \ldots, K$ from the multivariate normal posterior distribution $\mathrm{T}_{\mathrm{miss},i} \mid \mathrm{T}_{\mathrm{nearby}}$, and computing the variance between the samples and the posterior expectation:

$$\widehat{\mathrm{var}}(\mathrm{err}) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{var}_i \left\{ \mathrm{T}_{\mathrm{miss},i}^{(k)} - \mathbb{E}(\mathrm{T}_{\mathrm{miss},i} \mid \mathrm{T}_{\mathrm{nearby}}) \right\} \tag{23}$$

Similarly, to estimate the MSE expected by the model we use the MCMC draws $\tilde{\mathrm{T}}_{\mathrm{miss}}^{(k)}$, $k = 1, \ldots, K$ from SmoothHMC, and compute the MSE between the samples and the posterior expectation:

$$\widehat{\mathrm{MSE}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \left[\tilde{\mathrm{T}}_{\mathrm{miss},i}^{(k)} - \mathbb{E}(\mathrm{T}_{\mathrm{miss},i} \mid \mathrm{T}_{\mathrm{nearby}}, \mathrm{T}_x, \mathrm{T}_n) \right]^2 \tag{24}$$

When evaluating models, we want the errors to be small, and so the error variance and MSE to be low. A well-calibrated model should also have the estimated error variance (23) and MSE (24) close to their empirical values (22) and (21) respectively.

These diagnostics for our first spatio-temporal model, the product of squared exponentials, are found in the first row of Table 2. The error variance using only nearby measurements is already fairly low, with typical errors of order $\sqrt{1.59} = 1.26\,^{\circ}\mathrm{C}$. Incorporating T_n and T_x using SmoothHMC reduces it further to $\sqrt{1.12} = 1.06\,^{\circ}\mathrm{C}$. However, the model is overly optimistic, and the expected errors underestimate the empirical errors.

6 Improving the Basic Model

In this section, we develop more sophisticated Gaussian process covariances than the simple product of squared exponential kernels $\mathrm{k}_{\mathtt{SExSE}}$ (7). We then assess whether these models improve the variogram and the predictive diagnostics that we presented in Section 5.

The most salient feature of the empirical variogram that is not captured by the $\mathrm{k}_{\mathtt{SExSE}}$

covariance 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 processes allow for periodic components of the covariance, for example the periodic squared exponential covariance function, which we use with a 24-hour period

$$k_{24}(t,t') = \sigma_{24}^2 \exp\left[-\frac{2}{\ell_{24}^2}\sin^2\left(\pi\frac{t-t'}{24\text{ hrs}}\right)\right]. \tag{25}$$

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

$$\mathrm{k}_{\texttt{SESE}_{24}}(\boldsymbol{x},\boldsymbol{x}',t,t') = k_{\text{time}}(t,t')\cdot k_{\text{space}}(\boldsymbol{x},\boldsymbol{x}') + k_{24}(t,t')\cdot k_{\text{space24}}(\boldsymbol{x},\boldsymbol{x}') + k_{\mu}(\boldsymbol{x},\boldsymbol{x}')\,. \tag{26}$$

We also propose a more complex model, which breaks up k_{time} into short-term, medium-term and long-term correlation components:

$$\begin{aligned}
\mathrm{k}_{\texttt{sumprod}}(\boldsymbol{x},\boldsymbol{x}',t,t') = {} & k_{\text{time1}}(t,t')\cdot k_{\text{space1}}(\boldsymbol{x},\boldsymbol{x}') && \text{(short-term variation)}\\
& + k_{\text{time2}}(t,t')\cdot k_{\text{space2}}(\boldsymbol{x},\boldsymbol{x}') && \text{(medium-term variation)}\\
& + k_{\text{time3}}(t,t')\cdot k_{\text{space3}}(\boldsymbol{x},\boldsymbol{x}') && \text{(long-term variation)}\\
& + k_{24}(t,t')\cdot k_{\text{space24}}(\boldsymbol{x},\boldsymbol{x}') && \text{(diurnal cycle)}\\
& + k_{\mu}(\boldsymbol{x},\boldsymbol{x}') && \text{(station mean)}
\end{aligned} \tag{27}$$

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

$$k_{RQ}(t,t') = \sigma^2\left(1+\frac{(t-t')^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{28}$$

and is multiplied by a spatial decay component, specified as a squared exponential (1) with variance fixed at 1. Fitted covariance parameters for $\mathrm{k}_{\texttt{SESE}_{24}}$ and $\mathrm{k}_{\texttt{sumprod}}$ are found in Table 1.

We now have three competing Gaussian process models, with covariance functions $\mathrm{k}_{\texttt{SExSE}}$, $\mathrm{k}_{\texttt{SESE}_{24}}$, and $\mathrm{k}_{\texttt{sumprod}}$ respectively. We can compare them in four ways. Firstly, the variogram fit in Figure 8 is visibly improved by the introduction of the the diurnal component in $\mathrm{k}_{\texttt{SESE}_{24}}$,

and by the additional spatio-temporal correlation decay components in $k_{\texttt{sumprod}}$. Secondly, the marginal log-likelihood is the quantity maximized by the parameter fitting procedure in (8), with maximized values found in the second column of Table 2. The more complex models indeed yield a higher log-likelihood, promising a better model fit which should yield better predictions. Thirdly, 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 of Table 2, and shows more mixed results. The diurnal model $k_{\texttt{SESE}_{24}}$ performs slightly worse than the simple $k_{\texttt{SExSE}}$ model, and $k_{\texttt{sumprod}}$ only yields a small improvement. Fourthly, we compare the mean squared error specified in (21) for imputations at the test station incorporating T_n/T_x. Results in the fifth column also show more modest improvements for the more complex models. That said, with an expected MSE closer to its true value, $k_{\texttt{sumprod}}$ does give better estimates of its own inaccuracy.

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 (Stein, 2012). 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 $k_{\texttt{sumprod}}$ covariance (27). This is the primary output of our imputation method, and the results are promising. Firstly, just like in the toy example presented in Section 4.1, 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

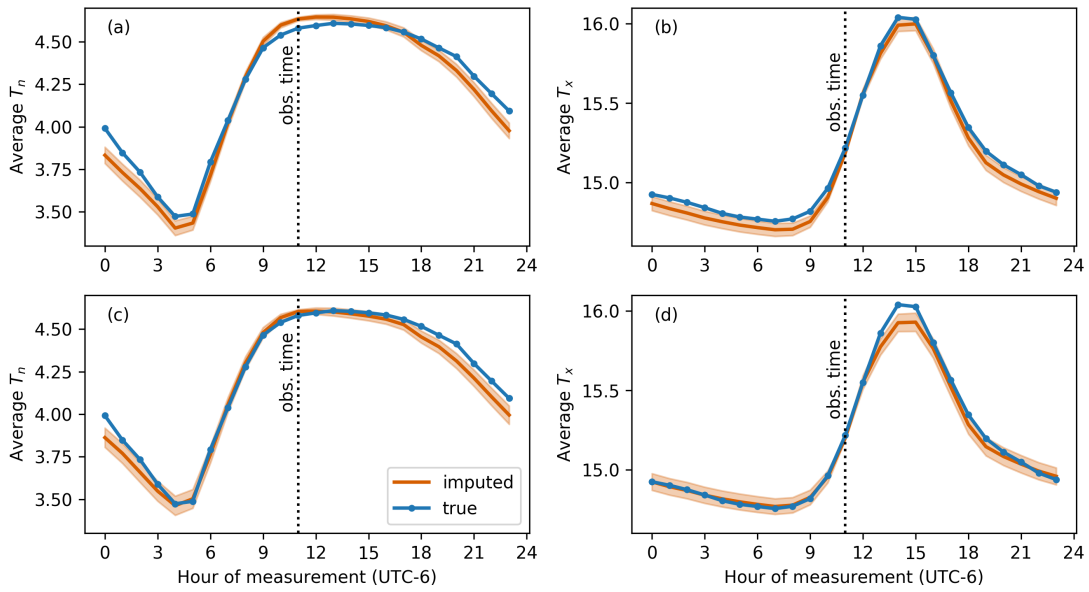

Figure 9: Average minimum (left) and maximum (right) daily temperature obtained under varying hour of measurement from KALO data (shown in blue), and from imputations of the withheld data (shown in orange with 2 SD envelope) obtained under the $k_{\texttt{SExSE}}$ covariance function (top) and the $k_{\texttt{sumprod}}$ covariance function (bottom).

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: during the May 31 measurement window (from May 30 at 11:00 to May 31 at 11:00) for example, 72.5% of SmoothHMC imputations reach their peak before 20:00 on May 30, 27.5% after 8:00 on May 31, and none in between. We view as a particular strength of our approach that the imputations are able to capture this ambiguity, rather than being restricted to a single mode of the posterior distribution.

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 3. Equipped with these imputations, is it possible to infer what the value of the summary statistic would have been for different measurement hours? This possibility is demonstrated in Figure 9, which shows the same summary statistic as in Figure 3 applied to the imputations as well as the (withheld) hourly data at KALO. It can be seen that the imputed summary statistics track within about 0.1 °C of the true

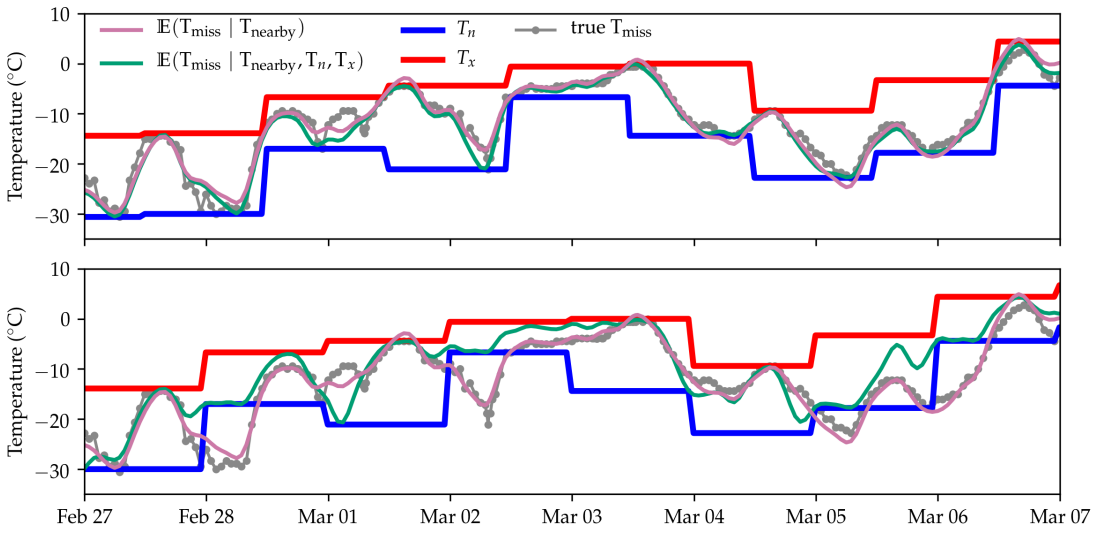

Figure 10: Constrained and unconstrained imputations in an eight-day window, assuming (top) the correct measurement hour (11:00 UTC-6), and (bottom) a wrong measurement hour (23:00 UTC-6). Assuming the wrong measurement time drives the constrained mean imputation away from the unconstrained mean imputation.

values. The product covariance $k_{\texttt{SExSE}}$ and the sum of products covariance $k_{\texttt{sumprod}}$ seem to perform equally well imputing the summary statistics for different times, but the $k_{\texttt{sumprod}}$ gives more honest, wider credible intervals.

8 Inference on Measurement Hour

Our analysis thus far has focused on the case where the hour of measurement `hr` is known in advance. This is a sometimes unrealistic assumption, and so inference on `hr` is desirable. It is conceptually straightforward to modify the measurement model (18) with a uniform prior on `hr`. However, `hr` affects which observations are attributed to each day's measurements, which has a discontinuous (observations suddenly jump from one day to the next) and non-differentiable effect on the posterior, and so Hamiltonian Monte Carlo becomes unviable. We therefore do not consider the introduction of a uniform prior on `hr` in Stan to be feasible.

Our procedure allows us to obtain imputation samples of T_{miss} conditional on T_{nearby}, T_n, T_x and `hr`. If we do so for $\texttt{hr} = 0, \ldots, 23$, is there information available in these samples

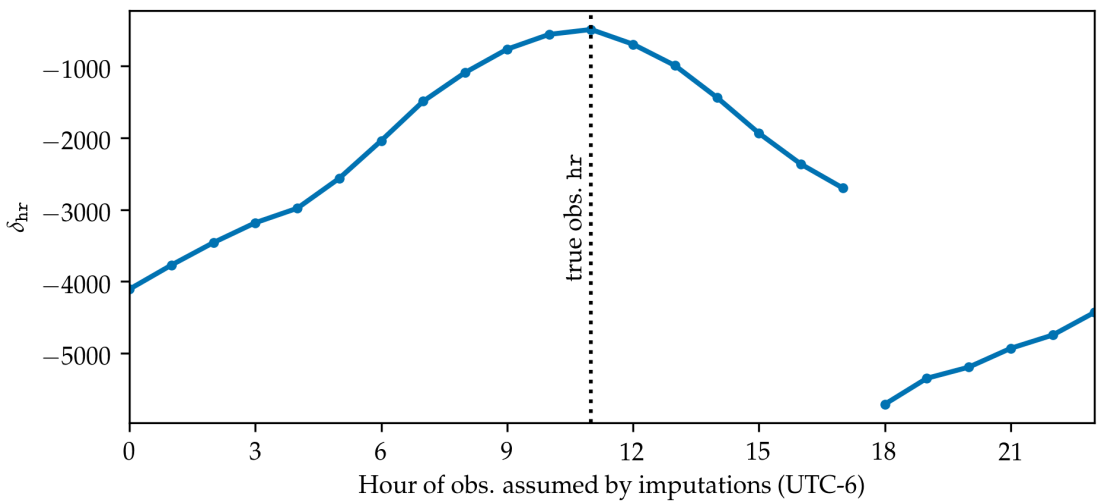

Figure 11: Concordance $\delta_{\mathtt{hr}}$ for imputations of temperatures at KALO assuming measurement hours $\mathtt{hr} = 1, \ldots, 24$. The true hour of measurement is 11:00, and obtains the highest $\delta_{\mathtt{hr}}$. Observations are associated with the date on which the observation occurs in the UTC timezone, which causes the discontinuity at 18:00 UTC-6.

to infer `hr`? We examine sample imputations in Figure 10 to gain intuition. Rather unsurprisingly, assuming an incorrect measurement time leads to wildly inaccurate imputations, for example on March 2nd. But notice also that assuming the wrong time 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 $\mathrm{T}_{\text{nearby}}$ and $\mathrm{T}_n/\mathrm{T}_x$, caused by assuming the wrong `hr`. We therefore propose to calculate the probability $\delta_{\mathtt{hr}}$ of the mean constrained imputation under the unconstrained posterior given by (9), which we interpret as a measure of concordance between $\mathrm{T}_{\text{nearby}}$ and $\mathrm{T}_n/\mathrm{T}_x$:

$$\delta_{\mathtt{hr}} = \log \mathbb{P}(\mathrm{T}_{\text{miss}} = \mu(\mathtt{hr}) \mid \mathrm{T}_{\text{nearby}}), \text{ where } \mu(\mathtt{hr}) = \mathbb{E}(\mathrm{T}_{\text{miss}} \mid \mathrm{T}_{\text{nearby}}, \mathrm{T}_n, \mathrm{T}_x, \mathtt{hr}), \tag{29}$$

Intuitively, $\delta_{\mathtt{hr}}$ will drop when the wrong `hr` is assumed, and we may be able to infer the true `hr` by maximizing $\delta_{\mathtt{hr}}$. In Figure 11, we demonstrate this method on the withheld KALO time series, which has been replaced by $\mathrm{T}_x/\mathrm{T}_n$ observations made at $\mathtt{hr} = 11$. We use SmoothHMC to impute the withheld data for all of 2015 under each possible measurement

hour $\texttt{hr} = 0, 1, \cdots, 23$. For each set of imputations, we compute the posterior mean $\mu(\texttt{hr})$ from the SmoothHMC samples, and the concordance $\delta_{\texttt{hr}}$ (29) (by necessity, modified to treat the center of each 73-day prediction window as an independent prediction). Pleasantly, the concordance is highest when the true hour of measurement is used so that, in this example at least, the correct hour of measurement would be infered.

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 (T_n) and maximum (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 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 KALO track the true temperatures, within a root mean square error of 1.02 °C. We view as particularly encouraging that the imputations successfully capture the uncertainty and sometimes bimodality in the 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 errors 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 KALO, we have demonstrated a good ability to recover this information (Figure 9).

The average T_x or T_n 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 estimated using imputed time series.

Lastly, we discussed the possibility of inferring the hour of measurement `hr`. We gave some intuition for maximizing the concordance (29) in order to infer `hr`, and a single example where this strategy is successful. While promising, we lack a theoretical justification for this approach. It remains to be seen whether our approach is generalizable and successful in other examples, and whether it can be placed on sound theoretical bases. Ideally we would wish to estimate the posterior probability $\mathbb{P}(\texttt{hr} \mid \mathrm{T}_{\text{nearby}}, \mathrm{T}_n, \mathrm{T}_x)$, for example by sampling from the joint posterior of $\mathrm{T}_{\text{miss}}, \texttt{hr} \mid \mathrm{T}_{\text{nearby}}, \mathrm{T}_n, \mathrm{T}_x$, but this is computationally difficult. Furthermore, it would be desirable not merely to infer the hour of measurement for an entire year, but to detect changepoints: days on which the measurement practice changed from one hour of measurement to another. We leave these improvements to inference of the measurement hour to future work.

References

Baker, D. G. (1975). Effect of observation time on mean temperature estimation. *Journal of Applied Meteorology 14*(4), 471–476.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker,

J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles 76*(1), 1–32.

Della-Marta, P. and H. Wanner (2006). A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate 19*(17), 4179–4197.

Ducré-Robitaille, J.-F., L. A. Vincent, and G. Boulet (2003). Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology 23*(9), 1087–1101.

Easterling, D. R., T. C. Peterson, and T. R. Karl (1996). On the development and use of homogenized climate datasets. *Journal of climate 9*(6), 1429–1434.

Karl, T. R., C. N. Williams Jr, P. J. Young, and W. M. Wendland (1986). 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.

Menne, M. J., I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston (2012). An overview of the Global Historical Climatology Network-Daily database. *Journal of Atmospheric and Oceanic Technology 29*(7), 897–910.

Menne, M. J. and C. N. Williams Jr (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate 22*(7), 1700–1717.

Menne, M. J., C. N. Williams Jr, and R. S. Vose (2009). The US Historical Climatology Network monthly temperature data, version 2. *Bulletin of the American Meteorological Society 90*(7), 993–1007.

Peterson, T. C., D. R. Easterling, T. R. Karl, P. Groisman, N. Nicholls, N. Plummer, S. Torok, I. Auer, R. Boehm, D. Gullett, et al. (1998). Homogeneity adjustments of in

situ atmospheric climate data: a review. *International Journal of Climatology 18*(13), 1493–1517.

Quinonero-Candela, J., C. E. Rasmussen, and C. K. Williams (2007). Approximation methods for gaussian process regression. *Large-scale kernel machines*, 203–224.

Sherman, M. (2011). *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

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

Vincent, L. A., X. L. Wang, E. J. Milewska, H. Wan, F. Yang, and V. Swail (2012). A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. *Journal of Geophysical Research: Atmospheres 117*(D18).

Appendix A Derivation of the analytic posterior for toy example

In this appendix we derive and compute the conditional distribution $F_{X|X_{\max},X_{\min}}$ for the toy example of Section 4.1. 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:

$$\mathbb{P}(X_i \leq x \mid X_{\max}, X_{\min}) = \begin{cases} 0 & \text{if } x < X_{\min}\,, \\ 1 & \text{if } x \geq X_{\max}\,, \\ \mathbb{P}_{i\bullet} + (1 - \mathbb{P}_{i\bullet} - \mathbb{P}_{\bullet i}) \left[\frac{F_i(x) - F_i(X_{\min})}{F_i(X_{\max}) - F_i(X_{\min})} \right] & \text{otherwise.} \end{cases} \tag{30}$$

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

$$f_i(X_{\min}) f_j(X_{\max}) \prod_{k \neq i,j}^{100} \left(F_k(X_{\max}) - F_k(X_{\min}) \right), \tag{31}$$

which we compute for all i, j and normalize to obtain the 100×100 matrix $\mathbb{P}$ of probabilities of each pair of element occupying the extremes. 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 $\mathbb{P}(X_i \leq x \mid X_{\max}, X_{\min})$ over a range of x. Figure 6(b) shows the analytical quantiles of $F_{X|X_{\max},X_{\min}}$ marginally for each X_i.