TemperatureImputations

Maxime Rischard

May 5, 2017

Contents

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, which 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 higher-resolution temperature data.

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 (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 (Menne et al., 2009). While these inhomogeneities have a small effect on, e.g., 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 applications 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 certain changes in measurement practices affect different percentiles of daily temperature in different ways. To address this issue, some homogenization methods have also employed percentile matching techniques, wherein the adjustment to a timeseries after a breakpoint is a function of percentile (Della-Marta and Wanner, 2006; Trewin, 2013).

Here, we focus primarily on addressing the time of observation bias, as well as its time trend, its because of its known impact on the distribution of daily maximum and minimum temperature (Tx and Tn) measurements. The bias exists because Tx and Tn 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 highest and lowest temperatures over the past 24 hours would typically be representative of the high and low during the prior day. For convenience, however, most observers record data at a daytime hour instead. As can be seen in Fig. XX, measurements recorded in the early morning may not properly register the low of the night before if it was unusually warm. Similarly, measurements recorded in the late afternoon may not properly register the high of the prior day if it was usually cool. In both cases, this will lead to a reduction in the variance of Tx and Tn distributions, but the effect will be greater at low (high) percentiles for Tx (Tn).

If the time of observation remained constant over time, the 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 Tx and Tn over time (Menne et al., 2009).

The goal of our approach is to infer the true T_n and T_x 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 are trying to recover the values of Tx and Tn that may have been overwritten due to measurement practices. Furthermore, by employing a Gaussian process framework and nearby stations with hourly data, we are able to simulate multiple realizations of temperature timeseries at each station, thereby providing estimates of uncertainty.

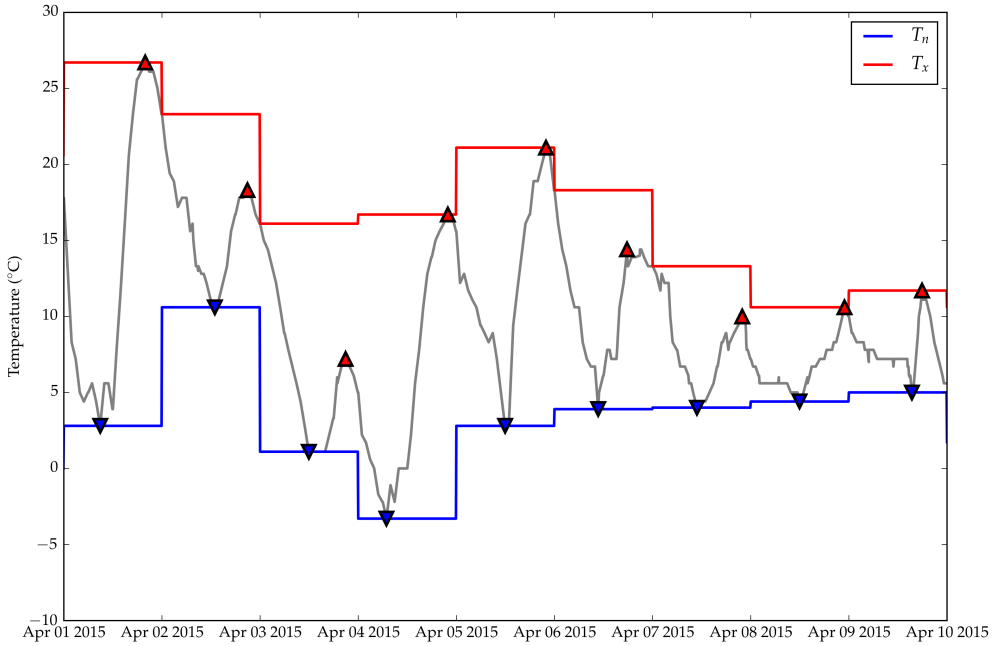
1.1 Illustration of bias induces by measurement hour

(Baker, 1975).

We illustrate the measurement 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. In Figure X, those ideal measurements are indicated by the red and blue triangles respectively. The actual measurements are obtained by dividing the data into 24 hour measurement windows, and extracting the minimum and maximum. For each window, we plot these extrema with a red and blue horizontal line.

On most days, the ideal measurement and the actual measurement coincide: the triangle is on that day's line. But there are also several misses. The most blatant example occurs on April 3rd, where the peak of the diurnal cycle is 7.2°C and occurs at 21:00 UTC. However, because the previous day was much warmer, 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.

This subtle bias in the daily records can in turn bias 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 the warmest time of day 21:00 UTC rather than 14:00 UTC (see Figure X). Conversely, the average T_n is colder by over 1°C if T_n is measured at 10:00 UTC (the coldest time of day on average) rather than 17:00 UTC.

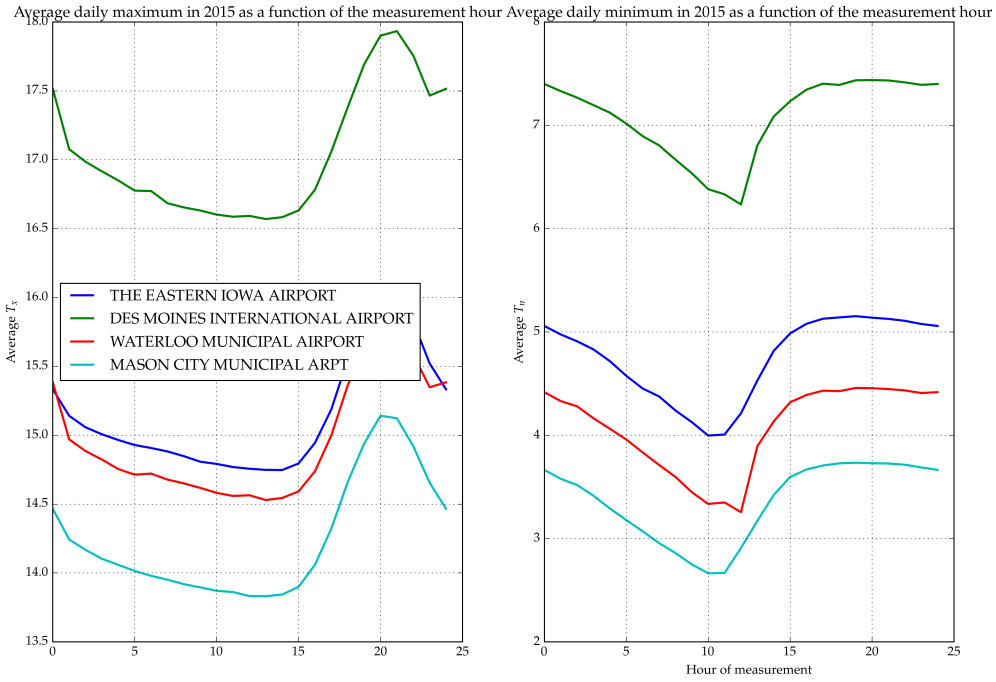


A climatologist studying weather variability might be interested in summary statistics such as the average absolute change in the daily temperature maxima and minima from one day to the next. The answer to that question too depends on the time of day at which the temperatures are recorded. Collecting 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 X, 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.

1.2 Proposed solution

We have seen that the daily maxima and minima do not faithfully record each diurnal cycle's peak and trough. The peaks on a relatively cold day can get overwritten by temperatures at either end of the measurement window that properly belong to the previous or the next diurnal cycle. Troughs on relatively warm days can be similarly overwritten. Our goal is to undo this damage, and recover estimates of summary statistics, such as the average daily maximum temperature, that do not suffer from the consequent bias. We need to address the erasure of information caused by the measurement mechanism, and therefore view this as a missing data problem.

Taking the missing data perspective, we seek to impute the hourly temperatures that have been replaced by a maximum and minimum over a 24 hour period. To do so, we use information from two sources: the recorded daily temperature extremes at the station of interest, and also hourly temperatures recorded at nearby meteorological stations. These hourly measurements are considered less reliable by climatologists, as they aren't as carefully documented, calibrated, and situated. The meterological stations are often in locations (like airports) where human activity will affect temperatures. 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 miscalibrated, the meterological data do contain valuable information about the hourly changes in temperatures on any given day. We therefore use them to inform the shape of the imputed temperature time-series at our location of interest, while we use the recorded temperature extrema to calibrate and constrain them.



2 First Spatiotemporal Model

To model measured temperatures at various locations and times, we use a spatio-temporal Gaussian process model. In its simplest form, we believe that temperatures from stations that are near each other are more correlated than distant stations, and that those correlations should also decay in time. In the spatial statistics literature, squared exponential covariance functions are commonly used to model correlations decaying as a function of distance. Ignoring the time dimension, we would model the simultaneous temperatures throughout a region as a Gaussian process, with the covariance of two locations $\mathbf{x}$ and $\mathbf{x}'$

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

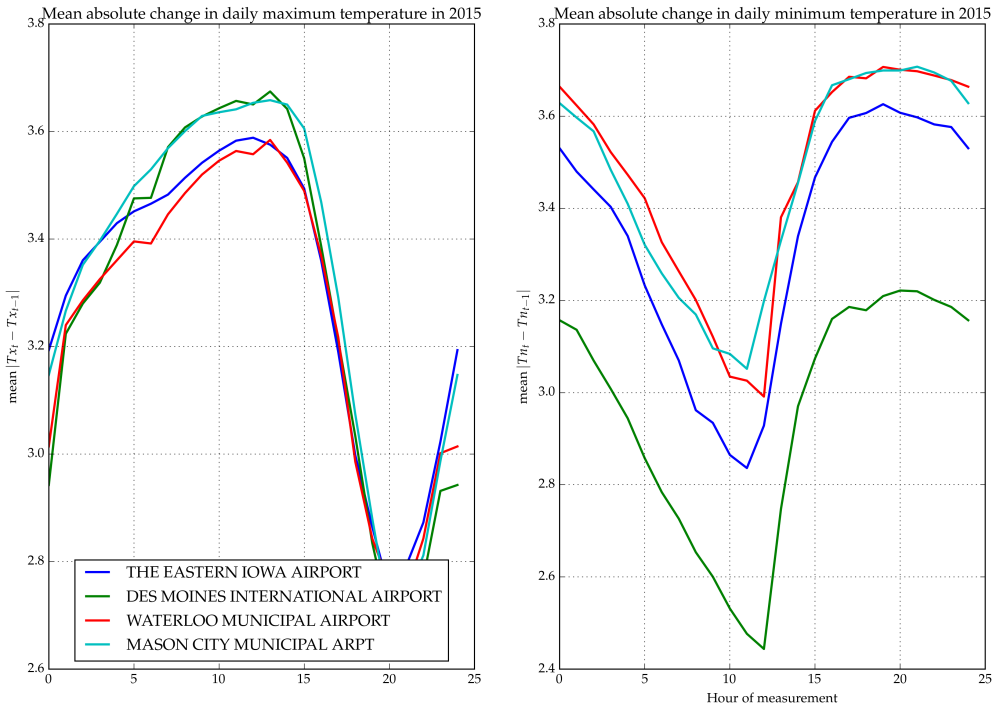
Similarly, ignoring the spatial dimension, the time series of temperatures at a single location can be modeled as a Gaussian process with covariance between two moments t and t'

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

We then combine the spatial and temporal model by multiplying the covariance functions

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

This gives us the covariance of the Gaussian process underlying the full spatio-temporal model of temperatures. To complete the model specification, we add a mean temperature for each station $\mu_{station[i]}$, and iid measurement noise ϵ_i.



$$T_i = \mu_{\text{station}[i]} + f(\mathbf{x}_i, t_i) + \epsilon_i \tag{4}$$

$$f(\mathbf{x}_i, t_i) \sim \mathcal{GP}\left(0, k_{st}(\mathbf{x}, \mathbf{x}', t, t')\right) \tag{5}$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right) \tag{6}$$

$$\tag{7}$$

3 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. However, the Iowa data includes 47,864 measurements, which is computationally infeasible to fit directly with a single Gaussian process. While approximation techniques exist to fit such large datasets, we chose the less efficient but simpler approach of dividing the data into 10-day chunks, modeled as independent Gaussian processes with shared hyperparameters. To simplify our implementation, we replaced the $\mu_{\text{station}}[i]$ terms by a spatial squared exponential component

$$k_\mu(\mathbf{x}, \mathbf{x}') = \sigma_\mu^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^\intercal(\mathbf{x}-\mathbf{x}')}{2\ell_\mu^2}\right) \tag{8}$$

with large variance σ_μ^2 and low lengthscale ℓ_μ added to the covariance function so that the spatiotemporal kernel becomes

$$k_{st}(\mathbf{x}, \mathbf{x}', t, t') = k_{time}(t, t') \cdot k_{space}(\mathbf{x}, \mathbf{x}') + k_\mu(\mathbf{x}, \mathbf{x}')\,. \tag{9}$$

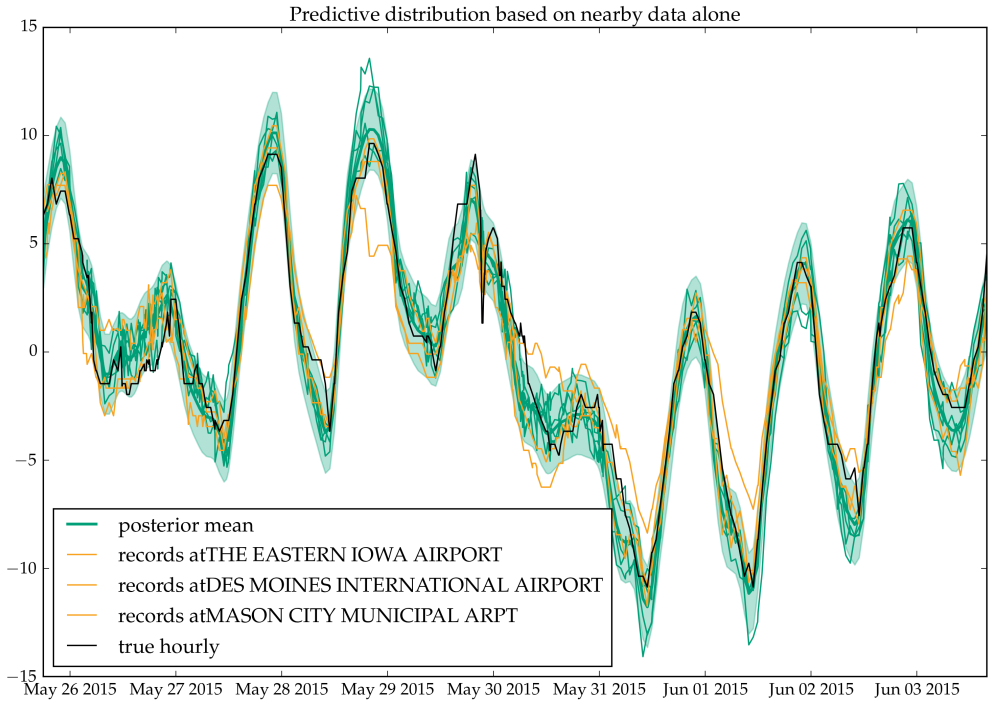


Figure 1: Predictive distribution using only nearby data and the simple product of square exponentials model. The orange lines are the measurements at nearby stations that are being used to inform the predictions. The black line is the true temperatures that have been withheld from the model, while the green line is the posterior mean of the predictions. The credible range (in green) is twice the standard deviations extracted from the diagonal entries of the posterior covariance matrix.

The model therefore has 4 free parameters: σ_{GP}, ℓ_t, ℓ_x and σ_ϵ. We optimize the marginal likelihood of the Iowa data as a function of these three parameters

$$\hat{\sigma_{GP}}, \hat{\ell}_t, \hat{\ell}_x, \hat{\sigma_\epsilon} = \underset{\sigma_{GP}, \ell_t, \ell_x, \sigma_\epsilon}{\arg\max} \left\{ \mathbb{P}\left(Y \mid \sigma_{GP}, \ell_t, \ell_x, \sigma_\epsilon\right) \right\}, \tag{10}$$

and obtain $\hat{\sigma_{GP}} = 3.73\,^\circ\,\mathrm{C}$, $\hat{\ell}_t = 2.7\,\mathrm{hours}$, $\hat{\ell}_x = 176.4\,\mathrm{km}$ and $\hat{\sigma_\epsilon} = 0.44\,^\circ\,\mathrm{C}$.

4 Imputations

Once we have a spatio-temporal Gaussian process model with optimized covariance parameters, we can use it to generate predictions at the station where we aim to generate imputations based on nearby measurements. Gaussian processes make this a closed-form procedure. We'll denote the temperatures we wish to impute as T_{miss} at times t_{miss} and location $\mathbf{x}_{miss}$ and those observed at nearby stations as T_{nearby}, at times t_{nearby} and locations X_{nearby}. Under the spatio-temporal model, T_{miss} and T_{nearby} are jointly multivariate normal, with mean zero and covariance given by $k_{st}(\mathbf{x}, \mathbf{x}', t, t')$. Standard results for conditioning within multivariate normals then yields

$$
\begin{aligned}
T_{miss} \mid T_{nearby} &\sim \mathcal{N}\left(\mu_{miss|nearby}, \Sigma_{miss|nearby}\right) , \\
\mu_{miss|nearby} &= \mathbb{E}\left(T_{miss} \mid T_{nearby}\right) \\
&= \operatorname{cov}\left(T_{miss}, T_{nearby}\right) \operatorname{cov}\left(T_{nearby}, T_{nearby}\right)^{-1} T_{nearby} , \\
\Sigma_{miss|nearby} &= \operatorname{var}\left(T_{miss} \mid T_{nearby}\right) \\
&= \operatorname{cov}\left(T_{miss}, T_{miss}\right) - \operatorname{cov}\left(T_{miss}, T_{nearby}\right) \operatorname{cov}\left(T_{nearby}, T_{nearby}\right)^{-1} \operatorname{cov}\left(T_{nearby}, T_{miss}\right) .
\end{aligned}
\tag{11}
$$

All covariance matrices can be obtained by plugging into k_{st}. For example, the ijth 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 X, we show an example of predictions obtained from this spatio-temporal model. We witheld measurements from the Waterloo Municipal Airport, and then used data from three nearby stations between May 25, 2015 and June 3, 2015 to predict the Waterloo temperatures during the same time window. This allows us to assess the quality of the predictions on this example.

Our aim, however, isn't just 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. To incorporate the additional information, we can use Bayes' theorem conditionally on T_{nearby}

$$
\mathbb{P}\left(T_{miss} \mid T_{nearby}, T_n, T_x\right) \propto \mathbb{P}\left(T_{miss} \mid T_{nearby}\right) \mathbb{P}\left(T_n, T_x \mid T_{miss}, T_{nearby}\right) \tag{12}
$$

so that the posterior using nearby measurements becomes the prior for the second stage of the analysis.

In fact T_n and T_x act as constraints on the predictions. For example, a T_x record means that in the 24 hours before the measurements, the temperature reached T_x but never exceeded it. This means T_n and T_x are deterministic functions of T_{miss}, and therefore

$$
\mathbb{P}\left(T_n, T_x \mid T_{miss}, T_{nearby}\right) = \mathbb{P}\left(T_n, T_x \mid T_{miss}\right) = \begin{cases} 1 & \text{if the constraint is satisfied,} \\ 0 & \text{otherwise.} \end{cases} \tag{13}
$$

Conceptually, we could therefore implement a valid imputation algorithm by drawing random samples from the posterior predictive multivariate normal distribution $T_{miss} \mid T_{nearby}$ obtained from nearby measurements, and only keeping the samples that satisfy this constraint. Unfortunately, the probability of a random draw exactly satisfying such a sharp constraint is zero, and so some tolerance would need to be introduced for overshooting or undershooting each day's T_n and T_x constraints. Imputing longer periods would then require either increasing the number of samples exponentially, or further loosening this tolerance. Ultimately, this rejection sampling strategy is therefore bound to fail.

Instead, we use the probabilistic programming language Stan to draw samples from the constrained predictive distribution. In Stan, we specify a probabilistic data-generating process for the observed temperatures, based on parameters and latent variables with accompanying priors. Stan then uses a Hamiltonian Monte Carlo (HMC) algorithm to draw sample from the posterior distribution for the parameters and latent variables. In our case, the observations are the daily maxima and minima, the only parameter is the average temperature at the station of interest μ_{miss}, and the latent variables are the missing unobserved hourly temperatures T_{miss}.

To summarize, the probabilistic model that we wish to draw posterior imputations of T_{miss} from is given in (14).

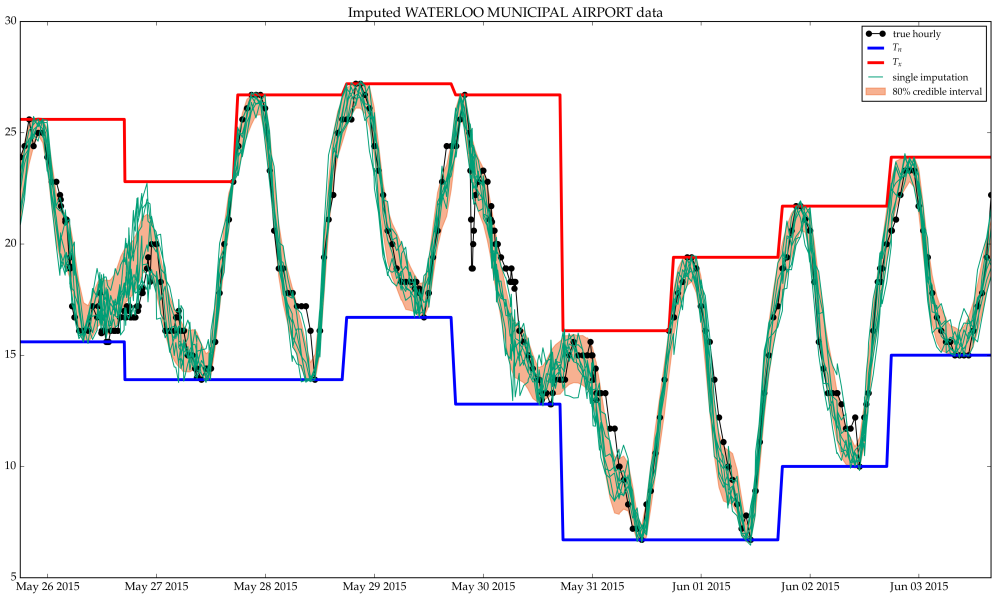


Figure 2: Imputations at Waterloo Airport from May 25, 2015 to June 3, 2015

$$
\begin{aligned}
\mu_{\text{miss}} &\sim \mathcal{N}(0, 100) && \text{(vague prior on mean temperature)} \\
f_{\text{miss}} &\sim \mathcal{N}\left(\mu_{\text{miss|nearby}}, \Sigma_{\text{miss|nearby}}\right) && \text{(posterior from } T_{\text{nearby}} \text{ becomes prior)} \\
T_{\text{miss}} &= \mu_{\text{miss}} + f_{\text{miss}} \\
T_x\,[\texttt{day}] &= \max_{i \in \{i\}_{\texttt{day}}} \{T_{\text{miss},i}\} && \text{(observe maximum in 24hr window)} \\
T_n\,[\texttt{day}] &= \min_{i \in \{i\}_{\texttt{day}}} \{T_{\text{miss},i}\} && \text{(observe minimum in 24hr window)} \\
\{i\}_{\texttt{day}} &= \left\{ i : \texttt{day} - 1 + \frac{\texttt{hour}}{24} < t_{\text{miss},i} \leqslant \texttt{day} + \frac{\texttt{hour}}{24} \right\} && \text{(indices of times in the 24hr window)}
\end{aligned}
\tag{14}
$$

The problem caused by the sharp constraint remains. At each step of the Markov chain, a proposal is made for T_{miss}. If the daily maxima and minima of this proposal coincide exactly with the measurements, then its posterior probability is finite. Otherwise it is exactly zero, with gradient also zero. HMC works by exploiting the gradient of the posterior, and therefore fails to converge in this situation, for reasons similar to the failure of the naive rejection sampler.

We can rescue the algorithm by replacing the max and min functions with the smoothmax and smoothmin functions, which take real inputs $x_1, \ldots, x_p$ and a sharpness parameter k and return

$$
\begin{aligned}
\text{smoothmax}\,(x_1, \ldots, x_p; k) &= \frac{1}{k} \log\left(\sum_{i=1}^{p} e^{k x_i}\right) \\
\text{smoothmin}\,(x_1, \ldots, x_p; k) &= -\,\text{smoothmax}\,(-x_1, \ldots, -x_p; k)
\end{aligned}
\tag{15}
$$

As $k \to \infty$, smoothmax becomes the maximum, and smoothmin becomes the minimum. Lower values of k give close approximations. When smoothmax replaces max and smoothmin replaces min, there is a small price in precision due to the approximation, but there is a huge computational benefit: gradients are now available. Furthermore, by adding a small amount of white noise with variance σ_ϵ to the T_x and T_n measurements, we ensure that the posterior doesn't go abruptly to zero when the conditions aren't met exactly. These modifications make HMC a viable algorithm to efficiently draw samples from the constrained

posterior. Setting k and σ_ϵ is a compromise between exactness and efficiency; we found $k = 10$ and $\sigma_\epsilon = 0.1$ to perform well. The modified model is given below, and approximates our ideal model (14).

$$\begin{aligned}
\mu_{\text{miss}} &\sim \mathcal{N}(0, 100) \\
f_{\text{miss}} &\sim \mathcal{N}\left(\mu_{\text{miss}|\text{nearby}}, \Sigma_{\text{miss}|\text{nearby}}\right) \\
T_{\text{miss}} &= \mu_{\text{miss}} + f_{\text{miss}} \\
T_x\,[\texttt{day}] &\sim \mathcal{N}\left(\underset{i \in \{i\}_{\text{day}}}{\text{smoothmax}}\{T_{\text{miss},i}; k = 10\}, 0.1^2\right) \\
T_n\,[\texttt{day}] &\sim \mathcal{N}\left(\underset{i \in \{i\}_{\text{day}}}{\text{smoothmin}}\{T_{\text{miss},i}; k = 10\}, 0.1^2\right)
\end{aligned} \tag{16}$$

Example imputations from this procedure are shown in Figure X. From May 25, 2015 to June 3, 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 "measurements" to constrain the imputed temperatures, and to infer the mean. Because we actually have hourly data for Waterloo, yet only fed our algorithm a reduction of this data to daily extremes, we can also plot the hidden temperatures, and see how faithfully the imputations reproduce them. We see that the imputations indeed track the true measurements very closely. The error bars satisfyingly narrow and widen in accordance to the amount of information available at each moments. On May 27th, we can see that the imputations capture the fact that the T_x record *could* have been set early in the measurement window, but more likely at its very end.

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

$$\gamma(\mathbf{h}, r) = \frac{1}{2}\mathbb{E}\left[(Y(\mathbf{x}, t) - Y(\mathbf{x} + \mathbf{h}, t + r))^2\right] = \text{var}(Y(\mathbf{x}, t)) - \text{cov}\left((Y(\mathbf{x}, t)), Y(\mathbf{x} + \mathbf{h}, t + r)\right). \tag{17}$$

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

$$\gamma(\mathbf{h}, r) = \sigma_\epsilon^2 + k(0, 0) - k(\mathbf{h}, r). \tag{18}$$

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 $\hat{\gamma}(0, r)$. For any pair of stations separated by $\mathbf{h}$ (fixed), we can also plot $\hat{\gamma}(\mathbf{h}, r)$. We then overlay the model's semi-variogram obtained through equation (18), resulting in Figure 3.

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

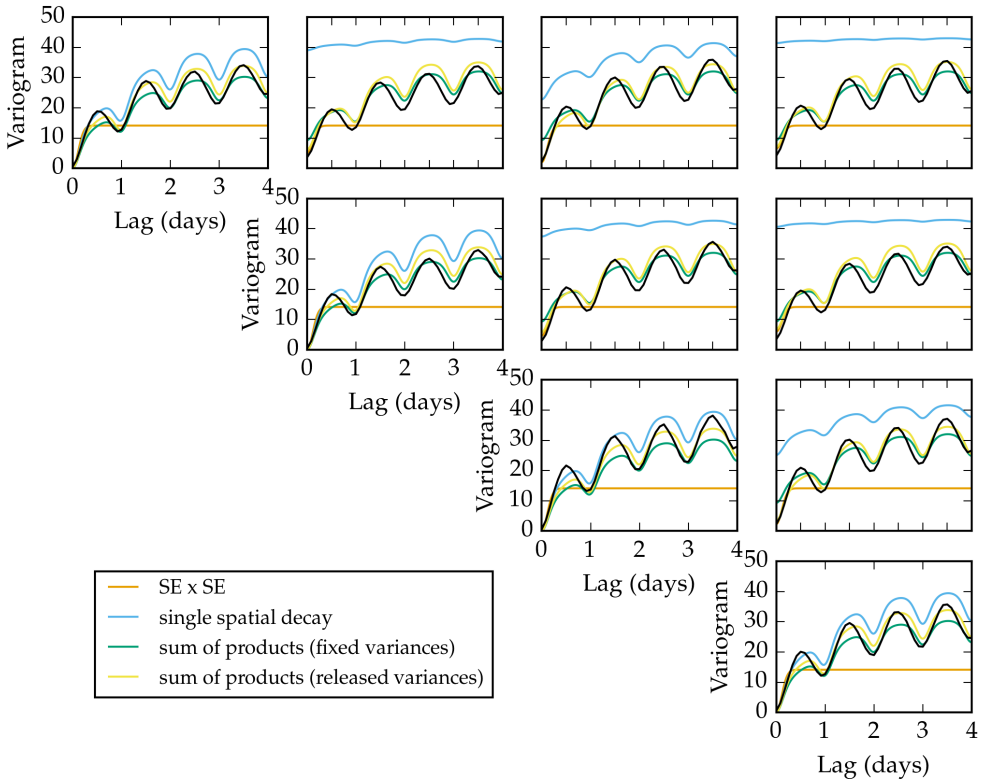


Figure 3: Semi-variogram

$$\mathrm{MSE}\left(\mathrm{err} \mid T_{\mathrm{nearby}}, T_x, T_n\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}\left(T_{\mathrm{miss},i} \mid T_{\mathrm{nearby}}, T_x, T_n\right) - T_{\mathrm{miss},i}\right]^2 . \tag{19}$$

This equation 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

$$\mathrm{var}\left(\mathrm{err} \mid T_{\mathrm{nearby}}\right) = \mathrm{var}_i\left\{\mathbb{E}\left(T_{\mathrm{miss},i} \mid T_{\mathrm{nearby}}\right) - T_{\mathrm{miss},i}\right\} . \tag{20}$$

Model	Log Likelihood	Var(err)	$\mathbb{E}$(Var(err))	MSE(err)	$\mathbb{E}$(MSE(err))
SE x SE	-55,614	1.589	0.875	1.104	0.614
SExSE + diurnal	-54,472	1.633	0.974	1.137	0.697
Sum of products, fixed variance	-48,589	4.991	8.791		
SoP, fixed temporal, free var	-47,082	1.314	2.321	1.150	0.897
SoP, completely free	-46,184	1.423	1.765	1.152	0.950
SoP, simpler	-45,945	1.319	1.190	1.069	0.823

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

$$\begin{aligned}
\mathbb{E}\left(\mathrm{var}\left(\mathrm{err} \mid T_{\mathrm{nearby}}\right)\right) &\approx \frac{1}{K}\sum_{k=1}^{K}\mathrm{var}_i\left\{T^{(k)}_{\mathrm{miss},i} - \mathbb{E}\left(T_{\mathrm{miss},i} \mid T_{\mathrm{nearby}}\right)\right\} \\
\mathbb{E}\left(\mathrm{MSE}\left(\mathrm{err} \mid T_{\mathrm{nearby}}, T_x, T_n\right)\right) &\approx \frac{1}{K}\sum_{k=1}^{K}\mathrm{MSE}_i\left\{T^{(k)}_{\mathrm{miss},i} - \mathbb{E}\left(T_{\mathrm{miss},i} \mid T_{\mathrm{nearby}}, T_x, T_n\right)\right\}
\end{aligned} \tag{21}$$

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 $\mathbb{E}\left(\mathrm{var}\left(\mathrm{err} \mid \cdot\right)\right)$ 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 X. The empirical error variance using only nearby measurements is already fairly low, with typical errors of order $\sqrt{1.589} = 1.26\,^\circ$ C. Incorporating the local measurements reduces it further to $\sqrt{1.104} = 1.05\,^\circ$ C. However, the model is overly optimistic, and the expected errors underestimate the true errors.

6 Improving 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 diagnostic measures that we developed in the previous sections.

The most salient feature of the empirical variogram that isn't captured by the `SExSE` model is the oscillation with a 24-hour period. It is intuitively obvious 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

$$k_{24}(t,t') = \sigma_{24}^2 \exp\left[-\frac{2}{\ell_{24}^2}\sin^2\left(\pi\frac{t-t'}{24\text{ hrs}}\right)\right]. \tag{22}$$

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)).

$$k_{SESE_24}(\boldsymbol{x},\boldsymbol{x}',t,t') = k_{time}(t,t')\cdot k_{space}(\boldsymbol{x},\boldsymbol{x}') + k_{24}(t,t')\cdot k_{space24}(\boldsymbol{x},\boldsymbol{x}') + k_{\mu}(\boldsymbol{x},\boldsymbol{x}'). \tag{23}$$

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.

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

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

$$k_{RQ}(t,t') = \sigma^2\left(1+\frac{(t-t')^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{25}$$

which is accompanied by its spatial decay kernel, specified as a squared exponential covariance. This more complicated kernel therefore has $3\times3\times2+2\times2 = 22$ 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 (10). The maximized log-likelihood can be found in the second column of Table XX, 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 (20) 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_{SESE_24} performs worse than the simple k_{SExSE} model, and $k_{sumprod}$ 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 (19) for predictions at the test station. Results in the fifth column show even more modest improvements for the more complex models.

We interpret these results as a reminder that predictions using Gaussian process are sensitive to model specification when extrapolating, but fairly insensitive to the model when interpolating [cite?]. Since our imputations interpolate the temperatures from nearby stations, further aided by the constraints imposed by the daily T_n and T_x measurements, the choice of model does not have a large impact on the performance of our procedure. This insensitivity can be seen as reassuring, as it (to an extent) reduces our need to worry about the incorrectness of our model.

7 Analysis

- show imputations on interesting days
- show imputations can capture two possible explanations for a measurement
- discuss possibility of inferring measurement time

8 Inference on measurement hour

Our analysis so 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 (16) 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_{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 4 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 (17 UTC), while the bottom plot is under an incorrect measurement time (5 UTC). 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 (11):

$$\begin{aligned} \mu(\texttt{hour}) &\equiv \mathbb{E}\left(T_{miss} \mid T_{nearby}, T_n, T_x, \texttt{hour}\right) \text{ (the mean imputed temperature)}, \\ \delta_{\texttt{hour}} &\equiv \mathbb{P}\left(T_{miss} = \mu(\texttt{hour}) \mid T_{nearby}\right). \end{aligned} \tag{26}$$

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 $\delta_{\texttt{hour}}$.

Fortuitously, our discrepancy measure $\delta_{\texttt{hour}}$ also admits a Bayesian interpretation: is is proportional to the marginal likelihood of `hour` under (admittedly fanciful) approximating assumptions. Ideally, we would evaluate the marginal likelihood $\mathbb{P}\left(T_n, T_x \mid T_{nearby}, \texttt{hour}\right)$, and then appeal to Bayes theorem to obtain a posterior on `hour`

$$\mathbb{P}\left(\texttt{hour} \mid T_n, T_x, T_{nearby}\right) \propto \mathbb{P}\left(T_n, T_x \mid T_{nearby}, \texttt{hour}\right) \mathbb{P}\left(\texttt{hour}\right). \tag{27}$$

However, marginal likelihoods are notoriously difficult to estimate from posterior samples [cite? Raftery 1994?]. The marginal likelihood is the normalizing constant for the posterior (12) of T_{miss}, and therefore for any T_{miss}

$$\mathbb{P}\left(T_n, T_x \mid T_{nearby}, \texttt{hour}\right) = \frac{\mathbb{P}\left(T_n, T_x \mid T_{miss}, \cancel{T_{nearby}}, \cancel{\texttt{hour}}\right) \mathbb{P}\left(T_{miss} \mid T_{nearby}, \cancel{\texttt{hour}}\right)}{\mathbb{P}\left(T_{miss} \mid T_n, T_x, T_{nearby}, \texttt{hour}\right)}. \tag{28}$$

The first term in the numerator is either one or zero, as discussed near equation (13). If we assume the constraint is satisfied, pick $T_{miss} = \mu(\texttt{hour})$, and assume that the posterior density evaluated at its mean does not depend heavily on the time of measurement, we obtain that the marginal likelihood is proportional to $\delta_{\texttt{hour}}$. Both assumptions are fanciful: the posterior mean generally will violate the constraint imposed by T_n and T_x, and therefore the likelihood $\mathbb{P}\left(T_n, T_x \mid T_{miss}\right)$ should in fact be zero. Furthermore, there is no reason to think the posterior density at the posterior mean does not depend on `hour`, but we might reasonably hope that the wrongness of this assumption does not overwhelm the signal contained in $\delta_{\texttt{hour}}$. This reasoning at least confirms that $\delta_{\texttt{hour}}$ captures information about the likelihood of `hour`, and that once renormalized it can be loosely interpreted as a posterior probability under a uniform prior.

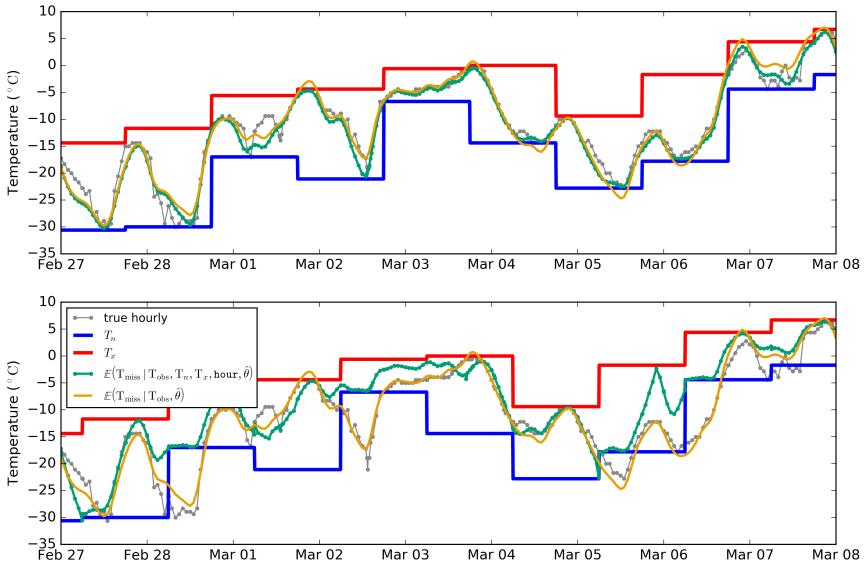


Figure 4: A sample window showing constrained and unconstrained imputations assuming (top) the correct measurement hour (17 UTC), and (bottom) the wrong measurement hour (5 UTC). Assuming the wrong measurement time drives the constrained mean imputation away from the unconstrained mean imputation.

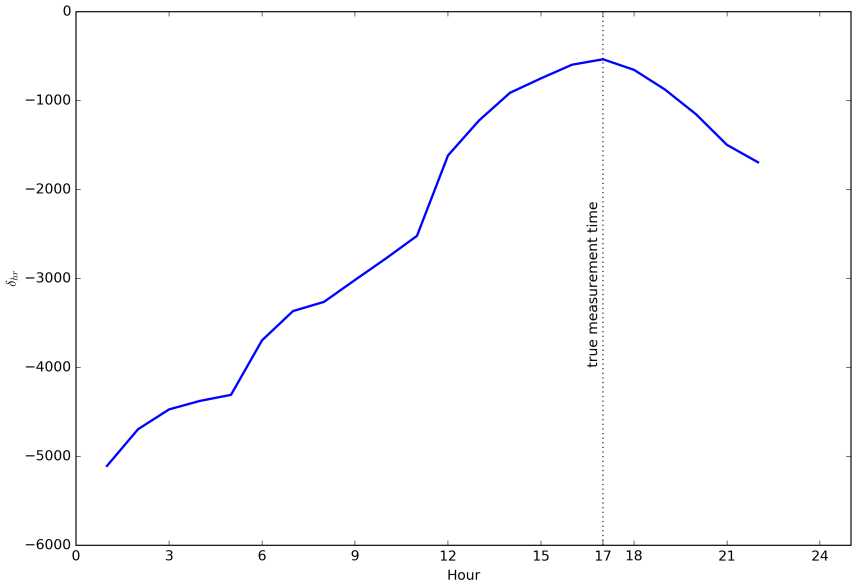


Figure 5: Discrepancy measure for imputations of temperatures at Waterloo Municipal Airport assuming measurement hours $\mathtt{hour} = 1, 2, \ldots, 24$. The true hour of measurement is 17, and obtains the highest $\delta_{\mathtt{hour}}$.

References

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

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., and Coauthors, 1998: Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology*, **18 (13)**, 1493–1517.

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)**.