Physically Constrained Spatiotemporal Modeling of Remotely Sensed Land Surface Temperature

Gavin Collins¹, Matthew Heaton¹, and Leiqiu Hu²

¹Department of Statistics, Brigham Young University, 223 TMCB, Provo, UT 84602 USA
²Department of Atmospheric Science, University of Alabama in Huntsville, NSSTC 3046, Huntsville, AL 35805 USA

Satellite remote-sensing is often used to collect important atmospheric and geophysical data that provide insight into spatial and temporal climate variability over large regions of the earth, at high spatial resolutions. Common issues surrounding such data include missing information in space due to cloud cover at the time of a satellite passing and large blocks of time for which measurements are not available due to the infrequent passing of polar-orbiting satellites. While many methods are available to predict missing data in space and time, in the case of land surface temperature (LST) data, these approaches generally ignore the temporal pattern called the "diurnal cycle" which physically constrains temperatures to peak in the early afternoon and reach a minimum at sunrise. In order to construct a complete, physically justifiable remotely sensed dataset, we parameterize the diurnal cycle into a functional form with unknown spatiotemporal parameters. Using multi-resolution basis functions, we estimate these unknown parameters from sparse satellite observations to obtain physically constrained estimations of LST. The methodology is demonstrated using two remote sensing datasets of LST in Houston, TX and Phoenix, AZ USA, collected by NASA's Aqua and Terra satellites.

Introduction

Satellite remote sensing technology is a common method of obtaining a wide variety of important data products. For example, land surface temperature (LST) is a remotely sensed variable that is of interest to scientists in multiple disciplines. Understanding the spatiotemporal patterns of LST in a region allows atmospheric and climate scientists to build more accurate climate models, identify urban heat islands, better understand atmospheric processes that may lead to severe weather events, and model concentrations of pollutants that are known to be highly correlated with LST (see ??, for details on these and other examples). The motivating examples for this research come from two remotely sensed datasets. The first contains measurements of LST in Houston (Harris County, Texas, USA) between June 28 and July 4, 2014, while the second contains LST measurements in Phoenix (Maricopa County, Arizona, USA) between September 11 and September 17, 2003. Both datasets were provided by the National Aeronautics and Space Administration (NASA), and were collected by the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on NASA’s polar-orbiting Terra and Aqua satellites.

As is typical of remotely sensed data products, MODIS collects LST on an aggregate scale at a high spatial resolution. The data are collected on relatively large grids of 1km by 1km areal units (i.e. pixels). The Houston grid is $151 \times 151$, yielding $151^2 = 22,801$ total areal units, while the Phoenix grid is $100 \times 150$, giving us a total of 15,000 areal units for that region. To give an idea of the spatial nature of the data, Figure ?? shows two examples of LST measurements for each of the two regions of interest. Note that, in the Houston region, there are many areal units for which no data were collected. This is partially due to the ocean in the southeast corner, as well as the inland lakes scattered throughout the region, where LST is masked because of the essential differences in surface material at these locations. These bodies of water are the main source of missing data in Figure ?? (a). However, there are many areal units, particularly in Figure ?? (b), for which data was not collected because of cloud cover at the time of the satellite passing. This cloud cover problem plagues a variety of remotely sensed data sets, especially in precipitous regions such as Houston, and inhibits researchers from using the data to perform scientific inference. In our dataset, 47.5% of the land data in Houston are missing due to cloud cover, while only 0.4% of the LST measurements are missing in Phoenix.

Both the Aqua and Terra satellites pass over each of the regions of interest twice daily, giving us a total of four potential LST measurements per day for each areal unit. With six days of data for each region, this means that there are $15,000 \times 6 \times 4 = 360,000$ potential temperature measurements in the Phoenix area, and $22,801 \times 6 \times 4 = 547,224$ potential measurements in Houston. To give a temporal view of the data, Figure ?? shows LST measurements in four areal units (two in Houston and two in Phoenix) for the entire time period over which the data were collected. Note that there are large blocks of time for which no data is collected due to the infrequent passing of the polar-orbiting satellites. Also note that there are times, principally for areal units in the Houston region, when data is missing at the time of a satellite passing due to cloud cover over the areal unit. As is common in the analysis of LST, we define a "day" as beginning and ending at the time of sunrise. Each day, the Terra satellite overpasses each region at around 11:00 and 22:00, while the Aqua satel-
The goal of this research is to develop statistical methodology that provides a temporally complete LST product for every areal unit within a specified region of interest. The first challenge in doing so is the high dimensionality of the data – in Houston, we have observed 243,346 total data points, and for Phoenix the observed dimension is 358,580 – which renders traditional spatial statistical methods computationally infeasible. As an additional challenge, LST is known to follow a physically constrained temporal pattern called the "diurnal cycle," which has a specific, physically justifiable functional form, stating that temperatures are maximized by early afternoon and coolest in the evening. Thus, an admissible LST data product would need to follow this diurnal cycle.

To demonstrate the necessity of explicitly incorporating the diurnal cycle into a statistical model, consider Figure 3 (a), which displays three days of LST data for one areal unit in the Houston region, along with a Gaussian process (GP) fit designed to interpolate LST. Utilizing a GP is a simple way to fit the data with a high degree of accuracy, but does not take into account scientific knowledge about the temporal patterns of LST. In contrast, Figure 3 (b) shows an estimation provided by explicitly incorporating a functional form of the diurnal cycle. Although the GP is able to more precisely model observed temperatures, the diurnal cycle provides more realistic estimations, even helping to validate and correct potentially inaccurate satellite readings. For example, it appears as though the Terra satellite measured LST around 16:00 on June 30 to be about 20°C, whereas the Aqua satellite measured LST about four hours later to be around 24°C. Given the time of day, this would be highly unusual, since LST is expected to decrease from around two hours after noon until sunset. The GP fits the data as well as it can, irrespective of this knowledge, while the diurnal cycle fit provides realistic estimates, while still accurately modeling the information obtained by both data sources.

The diurnal cycle provides a better fit than the GP in a few other ways as well. Notice that the GP fit is unrealistically jagged at times for which we have data, whereas the diurnal cycle is smooth, except at the knot points at the time of sunrise. The GP fits the data as well as it can, irrespective of this knowledge, while the diurnal cycle fit provides realistic estimates, while still accurately modeling the information obtained by both data sources.

To accomplish our goals, and thus provide scientists with a temporally realistic, and spatially and temporally complete data product, we propose to use novel statistical methods to spatially infill the diurnal cycle of LST for each of the 15,000
areal units in the Phoenix region, as well as the 17,930 land areal units in Houston. We employ a hierarchical Bayesian model where, in the first level of the hierarchy, we define a statistical model that is suitable for capturing the temporal pattern of LST in a single areal unit. This model incorporates an intuitive reparameterization of the diurnal cycle, and allows for interpolation in the presence of a temporally sparse dataset, as in the Houston and Phoenix data. To expand this methodology across many areal units, in the second level of the hierarchy, we use a set of spatial basis functions, designed to efficiently handle large quantities of spatial data, to model the parameters associated with the diurnal cycle. This methodology will produce a complete, physically constrained LST cycle for each areal unit in a general region of interest.

The remainder of the paper is outlined as follows. Section 2 (Methods) details our hierarchical statistical model, as well as the reparameterized functional form of the diurnal cycle. Section 3 (Results) applies this model to the Houston and Phoenix data, and compares the resulting LST predictions to those obtained by a temporal Gaussian process method. Finally, in Section 4 (Conclusion) we review our contributions, discuss potential limitations, and examine ideas to consider in future research.

Methods

A Spatially Constrained Spatiotemporal Model for LST

Let \( y_i(t) \) represent LST at time \( t \in \mathbb{R}^+ \) in areal unit \( i \in \{1,\ldots,n\} \), where \( n \) is the total number of areal units in the region of interest. We propose that

\[
y_i(t) \sim \mathcal{N}(\tau_i(t|\alpha_i, \mu_i), \sigma^2)
\]

where \( \tau_i(t|\alpha_i, \mu_i) \) is the functional form of the diurnal cycle of LST at location \( i \), and where \( \sigma^2 \) is the constant variance of \( y_i(t) \) across time for all areal units. Here we assume that \( \tau_i(\cdot) \) is an unbiased representation of the temporal pattern of LST, and that the residuals of the model are independent and normally distributed, with equal variance across time and throughout the spatial domain. For the Houston and Phoenix regions, we evaluated these assumptions via residual analysis.

Consider the diurnal cycle function \( \tau_i(\cdot) \) in Eq. (2). A common, accepted form of the diurnal cycle of LST originated from (2), who proposed that, for any given location on a particular day, LST is given by

\[
\tau(t) = \begin{cases} 
T_0 + T_a \cos \left( \frac{\pi}{2} (t - t_{max}) \right), & t < t_s \\
(T_0 + \delta_T) + \left[ T_a \cos \left( \frac{\pi}{2} (t_s - t_{max}) \right) - \delta_T \right] \left( \frac{t - t_s}{t_{max} - t_s} \right), & t \geq t_s
\end{cases}
\]

where \( t \in [0, 24] \) is measured in hours of hours after sunrise, \( T_0 \) is the residual temperature near sunrise, \( T_a \) is the temperature amplitude, \( t_{max} \) is the time of maximum temperature and \( t_s \) is the time at which the onset of nighttime attenuation occurs (i.e. the change point between the two equations in the system \( \tau(t) \)). The parameter \( \omega \) is the length of daylight hours, which is calculated via solar geometry, and is equal to \( \omega = \frac{4}{3} t_{max} \) (2). The parameter \( k \) is called the attenuation coefficient, and is defined as \( k = \frac{T_a \cos(u) - \delta T}{(T_a |\pi/\omega| \sin(u))} \) where \( u = \frac{\pi}{2} (t_s - t_{max}) \). Finally, \( \delta_T \) is implicitly defined as \( \delta_T = \tau(t \rightarrow \infty) - T_0 \). The five unknown parameters in this system include \( T_0, T_a, t_{max}, t_s, \) and \( \delta_T \), all of which are rather intuitive, with the exception of \( t_{max} \). For a more detailed explanation of their interpretation, see (2) and (3).

\[
\begin{align*}
T_a &= \frac{\mu_d - \alpha_d}{\cos(\pi/4) + 1} \\
T_0 &= \alpha_d + T_a \cos(\pi/4) \\
\delta_T &= \frac{T_a \cos(u)(T_a \cos(u) + T_0 - \alpha_d + 1) + \frac{\sin(u)(24 - t_s)(T_0 - \alpha_d + 1))}{24 - t_s - \cos(u)} - \cos(u)}{T_0 - \alpha_d + 1 - \frac{\sin(u)(24 - t_s)}{24 - t_s - \cos(u)}}
\end{align*}
\]

by which we may compute the diurnal cycle in terms of the more interpretable parameters \( \alpha_d, \alpha_{d+1}, \) and \( \mu_d \). To further simplify matters, in the application section of this paper we fix \( t_s \) (the start time of nighttime attenuation) at one hour before sunset, which (2) have shown to be a robust approximation of \( t_s \), incapable of introducing large error into the diurnal cycle model, and which substantially reduces the parameter space. We also fix \( t_{max} \) to be around two hours after solar noon, as suggested by (2), thus simplifying the diurnal cycle to three unknown parameters per areal unit per day. Because we only have a maximum of four measurements per areal unit per day (often fewer, especially for the Houston region), it would be difficult to justify estimating more parameters than this. Finally, in order to ensure continuity and differentiability throughout a diurnal cycle, we constrain the parameter space in three ways: \( \alpha_d < \mu_d, \mu_d > \tau(t_s|\alpha_d, \mu_d, \alpha_{d+1}) \), and \( \alpha_{d+1} \geq \tau(t_s|\alpha_d, \mu_d, \alpha_{d+1}) \).

As we have defined it, this form of the diurnal cycle is designed to model LST for a single day only, whereas we desire to build a model that is temporally continuous from day to day. To accomplish this, we adopt methods used by (2), and constrain temperature at the 24-hour mark on day \( d \) to be equal to temperature at the 0-hour mark on day \( d + 1 \). In other words, if \( \tau_d(t) \) is LST at the t-hour mark on day \( d \), then the constraint is \( \tau_d(24) = \tau_{d+1}(0) = \alpha_{d+1} \). Examining the systems of equations in Eq. (2) and Eq. (3), note that this
implies that the parameter $\alpha_{d+1}$ impacts LST after time $t_s$ on day $d$, and also during the entirety of day $d + 1$. Estimation of $\alpha_{d+1}$ is therefore impacted by LST measurements in the latter part of day $d$, as well as by all measurements on day $d + 1$, thereby tying measurements together across days. (The parameter $\mu_d$ is estimated using all measurements from day $d$, and only day $d$.). This constraint allows us to model LST over a period of $D$ days in a continuous fashion, while simultaneously reducing the parameter space for the entire cycle over the period to only $2D + 1$ parameters for each areal unit. Summarizing, our diurnal cycle function $\tau_d(t|\alpha, \mu)$ is valid for $D$ days of data, and is dependent on $2D + 1$ parameters, denoted by $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{i(D+1)})$ and $\mu_i = (\mu_{i1}, \ldots, \mu_{iD})'$, which are interpreted as follows. The parameter $\alpha_i$ for $d = 1, \ldots, D + 1$ represents LST at sunrise in location $i$ on day $d$, and $\mu_{i\tau}$ is the maximum LST at location $i$ on day $d = 1, \ldots, D$. Both $\alpha_i$ and $\mu_i$ are unknown parameters, because the satellite data do not include LST measurements at these times, and will be estimated from available data.

To intuitively explain our approach, Figure ?? (b) shows example diurnal cycles for various values of $\alpha_i$ and $\mu_i$ where $D = 1$ (thus there are 3 total parameters) and $t_{\text{max}}$ is the time of maximum temperature. Notice that this parameterization of the diurnal cycle is relatively flexible, allowing for different cycles depending on different parameter values. In this application we have elected to fix $t_{\text{max}}$ due to rather extreme temporal sparsity in our data but we note that this too could be considered an unknown parameter and estimated in more data rich settings. Finally, we employ the constraint $\tau_d(24) = \tau_{d+1}(0) = \alpha_{d+1}$ to obtain Figure ?? (c), which shows a diurnal cycle fit of the Phoenix satellite data displayed in Figure ?? (a), where $D = 3$.

Consider now, expanding the methodology in Eq. (??) across many areal units. We propose that the parameters of the diurnal cycle be subject to the spatial constraint

$$\Theta = XB$$

where $\Theta$ is an $n \times p$ matrix whose $i^{th}$ row contains the parameter vectors $\alpha_i$ and $\mu_i$ for areal unit $i$, $X$ is an $n \times q$ matrix of spatial basis functions, which will be detailed in the following subsection, and $B$ is a $q \times p$ matrix of coefficients for the $q$ basis functions corresponding to each of the $p$ elements of $\alpha_i$ and $\mu_i$. Intuitively, Equation Eq. (??) ties each of $\alpha_i$ and $\mu_i$ together spatially, allowing information to be borrowed across areal units to facilitate parameter estimation.

Modeling LST in this way allows us to estimate LST at an arbitrary areal unit $i^*$ within the general region of interest, at an arbitrary point in time $t^*$ within one of the $D$ days during which we have satellite measurements. Importantly, because of the spatial structure of the model, we need not have any data for areal unit $i^*$, and because of the temporal structure provided by the diurnal cycle, $t^*$ is potentially a time at which we have no temperature measurements. We will assign a prior distribution to the matrix $B$, and subsequently employ a Markov chain Monte Carlo (MCMC) algorithm to obtain posterior draws from $B$, which we will pre-multiply by $X_i^*$ (Equation Eq. (??)) to obtain draws of $\Theta_{i^*}$, checking at every step of the algorithm to ensure that each row in $\Theta$ satisfies the constraints on $\alpha_i$ and $\mu_i$, given above. This will allow us to obtain draws from the posterior distribution $y_{i^*}(t^*) \sim N(\tau_{i^*}(t^*|\alpha_{i^*}, \mu_{i^*}), \sigma^2)$, thus providing justifiable predictions of LST, along with proper estimates of uncertainty for any and all areal units, at any and all time points within the $D$ days for which we have data.

**Spatial Basis Functions for the Phoenix and Houston LST Data**

The method described in the previous subsection requires a set of areal basis functions $X$ suitable for describing the spatial variability of the diurnal cycle parameters over a general region of interest. While there are several choices of spatial basis functions to use (see ? for a review), we will adopt a multi-resolution areal basis function approach introduced by (? ). This approach builds on the methods of (? ), and is designed to model large areal datasets, such as those obtained in Phoenix and Houston, in a computationally feasible way. The approach was shown to be capable of accurately modeling the spatial structure of variables such as LST, even in the presence of large amounts of missing data, as in the case of the Houston dataset. We first describe these methods, and then apply them to the LST data.

(??) define a matrix of spatial basis functions $X$ to be the horizontal concatenation of the $q$ largest eigenvectors of the matrix $P^\perp A P^\perp$. Here, $A$ is an $n \times n$ adjacency matrix such that, if $a_{ij}$ is the $ij^{th}$ element of the $j^{th}$ column of $A$, then $a_{ij} = 1$ when areal units $i$ and $j$ are adjacent to each other, and $a_{ij} = 0$ otherwise. The $n \times n$ matrix $P^\perp$ is defined as $P^\perp = I_n - 1_n (1_n 1_n)^{-1} 1_n$, where $I_n$ is the $n \times n$ identity matrix, and $1_n$ is a vector of length $n$ where each element is 1. These basis functions were shown to appropriately model a variety of moderately-sized areal datasets.

The problem with using this approach for large data sets, such as the grid of temperatures in Houston and Phoenix, is that it quickly becomes infeasible to compute the eigenvectors of the $n \times n$ matrix $P^\perp A P^\perp$. (??) define a multi-resolution approach to overcome this issue. First they define low-resolution collections, each containing $k_r$ disjoint groups of areal units, where $r = 1, \ldots, R$ indexes the $R$ low-resolution collections, and where $k_r << n$ so that the eigendecomposition of $k_r \times k_r$ matrix $P^\perp A P^\perp$ associated with collection $r$ is feasible. Thus a number $q_r \leq k_r$ of basis functions may be obtained for each low-resolution col-
collection. Finally, \( \sum_{r=1}^{R} k_r \) high-resolution basis functions are then defined, where the groups in the high-resolution collections each contain only a single areal unit. Zeros are assigned to elements of a basis function corresponding to areal units not included in the collection, and these \( \sum_{r=1}^{R} (k_r) + R \) low- and high-resolution basis functions are horizontally concatenated to create a matrix \( X \) of multi-resolution areal basis functions.

To define the high- and low-resolution basis functions for this application, we follow the “shifted grid” practice outlined by (??), which was defined in order to avoid unnatural boundaries in predictive surfaces. This approach involves defining an initial low-resolution collection \( C_1 \), which contains \( k_1 \) disjoint groups of areal units. Using the methodology of (??), this collection then gives way to \( q_1 \leq k_1 \) basis functions, where \( q_1 \) is chosen by the researcher. To create high-resolution basis functions, these \( k_1 \) groups are then shifted one areal unit to the south, and one areal unit to the east, in order to form \( k_1 \) high-resolution collections \( H_1, \ldots, H_{k_1} \), which each give way to a single basis function.

To form additional low-resolution basis functions, the initial collection \( C_1 \) is then shifted two areal units to the south, and two areal units to the east, to form the low-resolution collection \( C_2 \), which contains \( k_2 \) disjoint groups, and which gives rise to \( q_2 \leq k_2 \) basis functions. As with collection \( C_1 \), collection \( C_2 \) is subsequently shifted one areal unit in the south-east direction to form \( k_2 \) additional high-resolution groups, which correspond to \( k_2 \) high-resolution basis functions. We repeat this process, stopping only when a new shifted grid would replicate an existing collection. If \( R \) low-resolution collections are formed, this procedure will lead to a total of \( \sum_{r=1}^{R} q_r \) low-resolution basis functions, and \( \sum_{r=1}^{R} k_r \) high-resolution basis functions.

Following suggestions of (??), in the Phoenix application we set \( k_1 = 280 \), with the largest group in the collection containing 64 areal units, but with some smaller groups. Using the shifted grid approach, this dictated that \( R = 4 \), with \( k_2 = 280, k_3 = 260 \), and \( k_4 = 247 \). We chose to set \( q_1 = q_2 = q_3 = q_4 = 100 \), giving us a total of \( q = 1,467 \) multi-resolution basis functions to describe the spatial variability of the diurnal cycle parameters over the 100 \( \times \) 150 region. These basis functions were horizontally concatenated to create an \( X \) matrix with \( n = 15,000 \) rows and \( q = 1,468 \) columns (including an intercept column).

For the 151 \( \times \) 151 Houston region, we set \( k_1 = 400 \), which led to \( R = 4 \), with \( k_2 = 400 \) and \( k_4 = 361 \). The largest group in any low-resolution collection again contained 64 areal units, but some groups were smaller. With the Houston grid, we also chose to set \( q_1 = q_2 = q_3 = q_4 = 100 \), giving us a total of 1,961 basis functions, but we were able to remove 156 of these because the non-zero elements in them corresponded only to areal units that were masked because they contained water. Thus the \( X \) matrix for Houston contained \( q = 1,961 - 156 + 1 = 1,806 \) columns.

Prior Distributions and Parameter Estimation

We chose to use a Bayesian approach to obtain estimates and uncertainty bounds for the unknown parameters in Equations Eq. (??) and Eq. (??), although maximum likelihood methods could also be employed. We now describe our choice of prior distributions for these unknown parameters.

The variance parameter \( \sigma^2 \) is assigned an inverse gamma prior \( \sigma^2 \sim \Gamma^{-1}(a, b) \), such that the posterior distribution of \( \sigma^2 \mid \{y, \Theta\} \) is conjugate inverse gamma. Equation Eq. (??) also contains two unknown diurnal parameter vectors \( \alpha_i \) and \( \mu_i \) for each areal unit. These parameters are contained in the matrix \( \Theta \), which is subject to the spatial constraint given by Equation Eq. (??), where the matrix of areal basis functions \( X \) is defined by the process described in the previous subsection. The coefficients of these areal basis functions are contained in the matrix \( B \), which is assigned a matrix-Gaussian prior, \( B \sim N_{n,p}(M, \Sigma, S) \). The prior distribution of \( \Theta \) (which defines distributions of \( \alpha_i \) and \( \mu_i \) for each areal unit) is then given implicitly by the hierarchy specified in Equation ??, along with the constraints given in Section 2.1.

Except for the first row, the mean matrix \( M \) was set at a “null hypothesis” prior of zero. In other words, if \( m_{ij} \) is the \( j^{th} \) row of \( M \), then \( m_{ij} = 0 \) for \( j \in \{2, \ldots, q\} \). The first row of the mean matrix, \( m_1 \), serves as the mean vector for the first row of \( B \), which is the vector of coefficients for the intercept column of \( X \). With the other rows set equal to zero, \( m_1 \) serves as the vector of prior means for \( \alpha_i \) and \( \mu_i \) for each areal unit. Hence, to specify \( m_1 \), we consulted past air temperature data (??) from June/July for the Houston area, and from September for Phoenix, which gave us a range of reasonable values for \( \mu \) and \( \sigma^2 \). For Houston, we set \( m_1 = (21, 32, 21, 32, \ldots, 21) \) (where \( m_1 \) is of length \( p = 2D + 1 = 13 \)), implying a prior belief that LST at the time of sunrise is around 21\(^\circ\) C, and that maximum LST is around 32\(^\circ\) C. For Phoenix, we set \( m_1 = (20, 52, 20, 52, \ldots, 20) \) (where \( m_1 \) is again of length \( p = 13 \)), reflecting the wide range of temperatures typical of a summer day in the American Southwest region. Note that these prior distributions ensure that each row of \( \Theta \) satisfies the constraints of the diurnal cycle, given in Section 2.2. It should also be noted that there are essential differences between air temperature and LST, which we took into account in selecting these prior distributions.

Because we lacked prior knowledge pertaining to the correlation structure of \( B \), we chose to set \( V = I_p \) and \( S = s^2 I_q \), where the matrices \( I_p \) and \( I_q \) are the \( p \times p \) and \( q \times q \) identity matrices, respectively. We chose \( s^2 = 1,000 \) as an arbitrary large number to reflect a relatively large degree of uncertainty about our choice of the mean matrix \( M \). Finally, we set \( t_{max} = 7.75 \) for the Phoenix area, and \( t_{max} = 7.95 \) for Houston, which corresponded to approximately 2 hours after solar noon, as suggested by (??).

With all prior distributions specified, we employed an MCMC algorithm, which we ran for 40,000 iterations, but we thinned the resulting chains by a factor of 4 to alleviate autocorrelation, and we dropped the first 5,000 draws, leaving us with 5,000 draws from the posterior distributio of each parameter. We monitored convergence of posterior chains...
with trace plots and effective sample size calculations, and found that the diurnal cycle parameters in $\Theta$ converged and mixed well. However, the posterior draws of $\sigma^2$ exhibited a downward trend across iterations for both the Phoenix and Houston data sets. We believe that this convergence issue is a result of the large amount of data that we are using to estimate $\sigma^2$. However, the range of the posterior draws of $\sigma^2$ was small in both cases ($[2.47, 2.53]$ for Houston and $[6.00, 6.08]$ for Phoenix). This, combined with the fact that $\Theta$, which contains the parameters we are most interested in, converged well, leads us to conclude that the lack of convergence in $\sigma^2$ has a negligible effect on overall inference. As will be shown in the next section, our methods allowed us to effectively accomplish our goal of infilling the entire diurnal cycle of LST for each areal unit in the Phoenix and Houston areas, while accounting for uncertainty in our estimations.

Results

Evaluating Model Assumptions

From Equation ??, note that our model makes a few basic assumptions about the variance structure of the residuals. First, we assume that all residuals (across time and space) are uncorrelated. Analysis of residuals across time, but within individual areal units, revealed little evidence of lag 1 autocorrelation, though we did find a surprising number of areal units exhibiting evidence of negative autocorrelation in the Houston region. We believe that this is because the diurnal cycle tends to split neighboring data points, i.e. it underestimates one while overestimating the next, which we consider to be a desirable property, as shown in Figure ??, at each satellite passing, residuals exhibited some fast-decaying spatial correlation, which we deemed negligible. Residual plots revealed moderate evidence of heteroskedasticity across time (more so for the Phoenix residuals, than for the Phoenix residuals), but we felt that the computational cost and model complexity required to treat $\sigma^2$ as a continuous variable across time would be unjustifiable, given that the heteroskedasticity was not extreme. To test this assumption more formally, we fit a model using a temporally continuous variance parameter, but found that the corresponding predictive accuracy was slightly lower, and that the coverage of the corresponding 95% prediction intervals was further from 95%, as compared to the model with a constant variance across time. Thus, we elected to retain the constant-variance model.

More concerning was the greater amount of evidence for heteroskedasticity across areal units in both the Houston and Phoenix regions. To evaluate the cost of modeling this spatial heteroskedasticity, we performed a cross-validation process using all six days of data for a relatively small subset of the spatial domain of the Phoenix region. We randomly formed a test set containing 10% of the LST measurements in the subset and fit a model using the remaining 90%. We then used the model to predict LST at the test locations and times, and evaluated the performance of the predictions via root mean-squared error (RMSE). We performed the same cross-validation process with a model that allowed $\sigma^2$ to vary freely across areal units, and the result was a slightly lower degree of predictive accuracy. Thus we concluded that a single $\sigma^2$ across the spatial domain would provide a more accurate model.

Model Fit

As previously mentioned, parameters for the diurnal cycle parameter are interpretable, and are of interest in of themselves. Figure ?? shows estimates and 95% credible intervals for the temperature at time of sunrise on day 2 (June 30, 2014), i.e. $\hat{\alpha}_2$ for the Houston region. These estimates alone could be used for scientific inference, and we will also use them to obtain estimates of the entire diurnal cycle for each of the areal units in the Houston region. Similarly, Figure ?? shows estimates and 95% credible intervals for maximum LST on day 5 (September 15, 2003), i.e. $\hat{\mu}_5$ for the Phoenix region. Notably, the credible intervals are much narrower for the Phoenix region, which we believe to be a result of fewer missing observations in this data set. Table ?? evaluates in-sample predictions for the model applied to the Houston data. The root mean square error (RMSE) for time $t$, at which a satellite passing occurred, is given by $\text{RMSE}_t = \sqrt{\frac{1}{5} \sum_{i=1}^{20} (\hat{\gamma}_i(t) - \hat{\gamma}_i(t))^2}$, where $\hat{\gamma}_i(t)$ is the predicted LST for areal unit $i$ at time $t$. RMSE is above $3^\circ$ C for satellite passings 11, 18, and 21, but average RMSE is quite low, at $1.56^\circ$ C. There is substantial bias for some of the satellite passings (passings 11, 18, and 21 again stand out, with bias greater than $3^\circ$ C) but bias is expected at some points in time, due to the constraint of the diurnal cycle. As shown in Figure ??, this bias is desired in some instances.
Furthermore, average bias is zero, when rounded to two decimal places. An evaluation of in-sample coverage and prediction interval width is also provided for each satellite passing. Coverage is above 95% for 17 of the 24 satellite passings, but coverage is extremely low for passings 11, 18, and 21, where there is a large amount of bias, which brings down overall coverage to 90.3%. The average interval width is below 7°C for 15 of the satellite passings, but interval widths tend to be wider for satellite passings early and late in the week. We conjecture that this is a result of fewer data points feeding into the estimation of LST at these times. Finally, Table ?? lists the percent of missing data at each satellite passing.

Table ?? is analogous to Table ??, but it evaluates in-sample predictions for the Phoenix data. Here, RMSE is above 3.5°C for satellite passings 5, 13, and 21, but average RMSE is quite low, at 2.22°C. Passings 5, 13, and 21 each have a bias above 3°C in absolute value but this is again expected. As with the Houston data, average bias is zero when rounded to two decimal places. In-sample prediction interval coverage is above 95% for 21 of the 24 satellite passings, and below 87% for only the 5th passing, where the greatest amount of predictive bias occurs. Overall coverage is above 95%. The average interval width is very consistent across time: the smallest average interval width is 9.73°C, and the largest is 9.92°C. The fact that there are larger errors and interval widths in the Phoenix data is unconcerning, because the range of the data is much greater than in the Houston data set.

Figure ?? provides a visualization of our model’s predictive ability for a single satellite passing over Houston. Panel (a) shows actual satellite readings of LST just an hour before the time of maximum temperature on July 1, 2014. Model estimates are shown in panel (b), and lower and upper bounds of the 95% prediction intervals are shown in panels (c) and (d), respectively. From row 14 of Table ??, we see that the RMSE of these estimates is 1.34°C, and that the coverage of the interval is 97.3%. For this particular satellite passing, bias is negligible at −0.08°C. Figure ?? is analogous to Figure ??, but for Phoenix data in the early morning hours of September 15, 2003. From row 16 of Table ??, we see that the RMSE of these estimates is 2.03°C, and that bias is relatively large, at −1.73°C. Coverage for the associated in-sample prediction intervals is 99.7%, indicating that these intervals may be wider than necessary.

A temporal view of model predictions is shown in Figure ??, where satellite-observed LST is depicted with blue dots, continuous predicted values are shown with solid black lines, and 95% prediction intervals are shown with dashed black lines for two areal units in each region. Notably, Figure ?? (a) illustrates that we are able to obtain estimates of LST for even a day when no satellite measurement was recorded for an areal unit, because of the spatial constraint on the diurnal cycle parameters. Also note that the prediction intervals cover most data points for these areal units – one exception is in panel (b), where the 3rd data point on June 30 is below the confi-
Showing measurements of LST at 12:42 on June 30, as collected from the satellite at this point in time. However, this data point is unusual, and we are slow to believe that LST was accurately measured by the satellite. The first row of panels illustrates a major benefit of directly incorporating the diurnal cycle into our LST model. The first row of panels shows measurements of LST at 12:42 on June 30, as collected by the Aqua satellite, and at 23:00 on June 30, as collected by the Terra satellite. The second row shows a fit obtained by applying a temporal Gaussian process model to the coefficients associated with spatial basis functions, which are associated directly with the LST data via linear regression (see Equation 2, for the details of this model). It is evident from this figure that the Gaussian process model fits the data quite well. However, by the time of maximum temperature (around 14:00) the Gaussian process predictions are substantially lower than the observed data, but maximum and minimum temperatures occur too early in the day, as compared to the more realistic diurnal cycle fit. In contrast, the diurnal cycle model yields more realistic temporal predictions, with the highest LST predictions occurring at the time of maximum temperature, and the lowest occurring at the time of sunrise. We argue that the diurnal cycle fit is a more accurate representation of LST data, and will provide scientists with a more useful inference tool.

**Conclusion**

In summary, we successfully defined a simple multi-day reparameterization of the diurnal cycle of LST, with highly interpretable parameters. We then generated a parsimonious
statistical model using the diurnal cycle, and applied this model to two large areal data sets, employing a spatial constraint on the diurnal cycle parameters to account for the spatial variability of LST. The resulting predictions were reasonable from an applied prospective, with estimates reaching a maximum around two hours after solar noon, and a minimum at the time of sunrise. Predictions were accurate, and their overall bias was zero.

As shown in Tables ?? and ??, our model imperfectly quantifies the uncertainty associated with our predictions. We looked into more properly accounting for this uncertainty by involving a continuously defined variance parameter, but this attempt actually made the problem somewhat worse; this may be a result of sparse data in the temporal dimension, and thus we may not have enough data to estimate a continuous variance function. We also built a model in which the variance was allowed to vary freely across space, but this too proved to be less effective than a single variance parameter, perhaps because the high dimensionality of the spatial domain led to such a large number of variance parameters.

On the whole, our methods provide a practical and accurate way to infill the complete diurnal cycle of LST across multiple days for large areal data sets collected by remote sensing satellites over a general region of interest.

Bibliography