Received 27 March 2023
Revised 25 September 2023
26 September 2023
Accepted 27 September 2023

Lightning forecast from chaotic and incomplete time series using wavelet de-noising and spatiotemporal kriging

Jared Nystrom, Raymond R. Hill, Andrew Geyer and Joseph J. Pignatiello
Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA, and
Eric Chicken
Department of Statistics, Florida State University, Tallahassee, Florida, USA

Abstract

Purpose – Present a method to impute missing data from a chaotic time series, in this case lightning prediction data, and then use that completed dataset to create lightning prediction forecasts.

Design/methodology/approach – Using the technique of spatiotemporal kriging to estimate data that is autocorrelated but in space and time. Using the estimated data in an imputation methodology completes a dataset used in lightning prediction.

Findings – The techniques provided prove robust to the chaotic nature of the data, and the resulting time series displays evidence of smoothing while also preserving the signal of interest for lightning prediction.

Research limitations/implications – The research is limited to the data collected in support of weather prediction work through the 45th Weather Squadron of the United States Air Force.

Practical implications – These methods are important due to the increasing reliance on sensor systems. These systems often provide incomplete and chaotic data, which must be used despite collection limitations. This work establishes a viable data imputation methodology.

Social implications – Improved lightning prediction, as with any improved prediction methods for natural weather events, can save lives and resources due to timely, cautious behaviors as a result of the predictions.

Originality/value – Based on the authors' knowledge, this is a novel application of these imputation methods and the forecasting methods.

Keywords Forecasting, Imputation, Wavelets, Kriging

Paper type Research paper

1. Introduction

Forecasters develop a risk assessment of lightning activity at Kennedy Space Center and Cape Canaveral Space Force Station (KSC/CCSFS) using a dense array of electric field mill (EFM) sensors. These sensors measure the ground-level electric potential within the atmosphere directly overhead each sensor, indicating changes in electromagnetic energy. These changes are phenomena shown to be predictive of future lightning activity (Aranguren *et al.*, 2012; Lopez *et al.*, 2012). The EFM network records data at 50 Hz, resulting in very large

This research is partially funded by the Omar Nelson Bradley Fellowships.

Data availability: The data associated with this paper were provided by the 45th Weather Squadron at Patrick Air Force Base and could be obtained by a request through that organization.

Journal of Defense Analytics and Logistics
Vol. 7 No. 2, 2023
pp. 90-102
Emerald Publishing Limited
2399-6439
DOI 10.1108/JDAL-03-2023-0003

data structures that are of high frequency and high volume. Furthermore, these datasets are autocorrelated in regards to both temporal timestamps and spatial distancing of the fixed EFM sensor sites.

While probabilistic methods and time series models have been used to try to exploit the EFM data for lightning prediction (Nystrom, 2021), recently machine learning approaches have been found promising. Hill (2018), Speranza (2019) and Cheng (2020) examined neural network approaches and achieved good prediction accuracy. Nystrom *et al.* (2021) used wavelet methods and further improved the prediction accuracy over the machine learning methods. However, each of these efforts was plagued by missing data issues.

As it is common across many types of sensors, the EFM sites periodically experience periods of time missing measurements. This can be due to routine site maintenance, sensor malfunction or a purposeful shutdown due to local disturbances that would perturb the sensor readings. These gaps in collection prove to be problematic in some machine learning and artificial intelligence applications as some methods are not robust to periods of missing data and will fail to learn properly. Deletions of incomplete records or imputation methods are used to preprocess the data for the machine learning algorithm. This study applies imputation methods on the spatially-based EFM time series, making use of the inherent autocorrelation structure in the data, resulting in improved modeling using machine learning and artificial intelligence techniques.

Imputation is a data preprocessing method which substitutes missing entries with estimated values. There are many imputation methods available based upon data type and application. The simplest imputation methods use a representative value for all missing entries, such as the mean, median or mode of available data. Time series imputation is a sub-discipline which takes into account the autocorrelation between timestamped values. For instance, use of time stamped observations of air pollutants to produce an estimate for missing values (Junger and De Leon, 2015). Autocorrelation in time series is the dependence of values between time-stamped observations. This results in a great deal of redundancy of the information within time series data, and if not accounted for can result in a model that overstates fit (Eshel, 2012). Time series imputation approaches include use of moving averages, extension of nearest observation, Kalman smoothing and linear or spline interpolation (Moritz and Bartz-Beielstein, 2017). Likewise, spatial imputation methods are a subdiscipline that estimates missing data values while accounting for autocorrelation present between spatially correlated measurements: For instance, the estimate of tree density measurements from nearby measurement sites within an especially dense forest (Robinson and Hamann, 2011).

This paper employs a spatiotemporal imputation technique that simultaneously accounts for autocorrelation between spatially correlated measurements collected as a time series. Wavelet methods are used as an additional preprocessing step, serving to de-noise the chaotic EFM measurements to allow faster convergence and estimation of spatiotemporal models. Instead of a purely time series or spatial model, spacetime approaches use all available data to infer predicted values. These methods prove to be highly useful in situations in which large amounts of a particular time series are missing and need to be estimated. Although complex in application, such methods are of increasing importance due to the increasing prevalence of modern sensor systems. Section 2 provides an overview of the EFM dataset, wavelet methods for de-noising a time series and spatiotemporal modeling techniques. Section 3 presents the methodology and results of wavelet techniques and spatiotemporal modeling as an imputation method. Section 4 applies the EFM dataset, to include values estimated by spatiotemporal kriging, using an existing methodology and compared to a baseline imputation method. Conclusions and applications for future research are provided in Section 5.

2. Methodology

2.1 EFM sensor network

Lightning activity is particularly concentrated in the KSC/CCSFS region of central Florida, as can be seen in the heat-map of Figure 1. Accurate and timely forecasts of lightning activity are essential to inform operational risk assessments that guide both flight line and space launch activities. Current studies indicate EFM networks can be predictive of lightning activity through either a relatively sudden change of polarity or an increase in magnitude of the atmospheric electric potential (Aranguren *et al.*, 2012; Lopez *et al.*, 2012). However, constant movement and churning actions within the atmosphere result in a chaotic response of electrostatic potential by the EFM network (Krider, 1989). Figure 2 provides three examples of typical and chaotic EFM measurements prior to observed lightning within KSC/CCSFS. Current literature also indicates a diurnal cycle to the EFM network at KSC/CCSFS

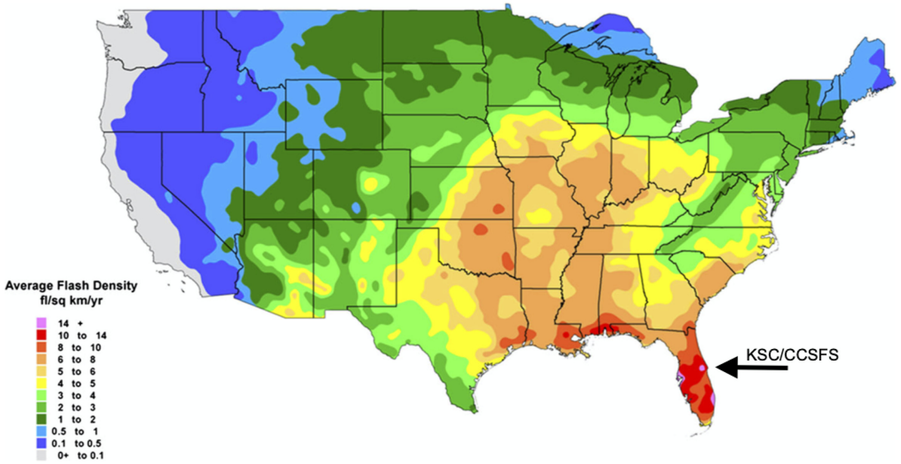


Figure 1.
Cloud-to-ground lightning flash density (1997–2010) for the USA from the national lightning detection network

Source(s): Figure courtesy of Roeder *et al*. (2015)

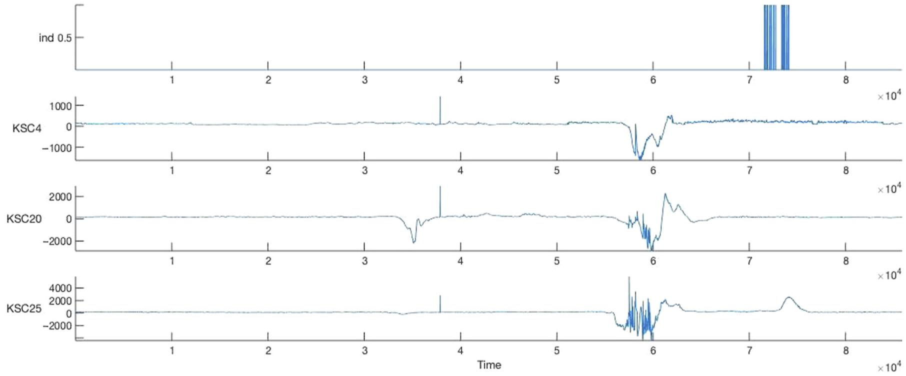


Figure 2.
Top subplot is binary response of observed lightning, followed by three typical EFM measurements chosen randomly across the entire KSC/CCSFS region over time in seconds

Note(s): The EFM measurements indicate a natural steady state in the absence of lightning, becoming increasingly chaotic as electromagnetic potential builds within the atmosphere

Source(s): Figure courtesy of authors

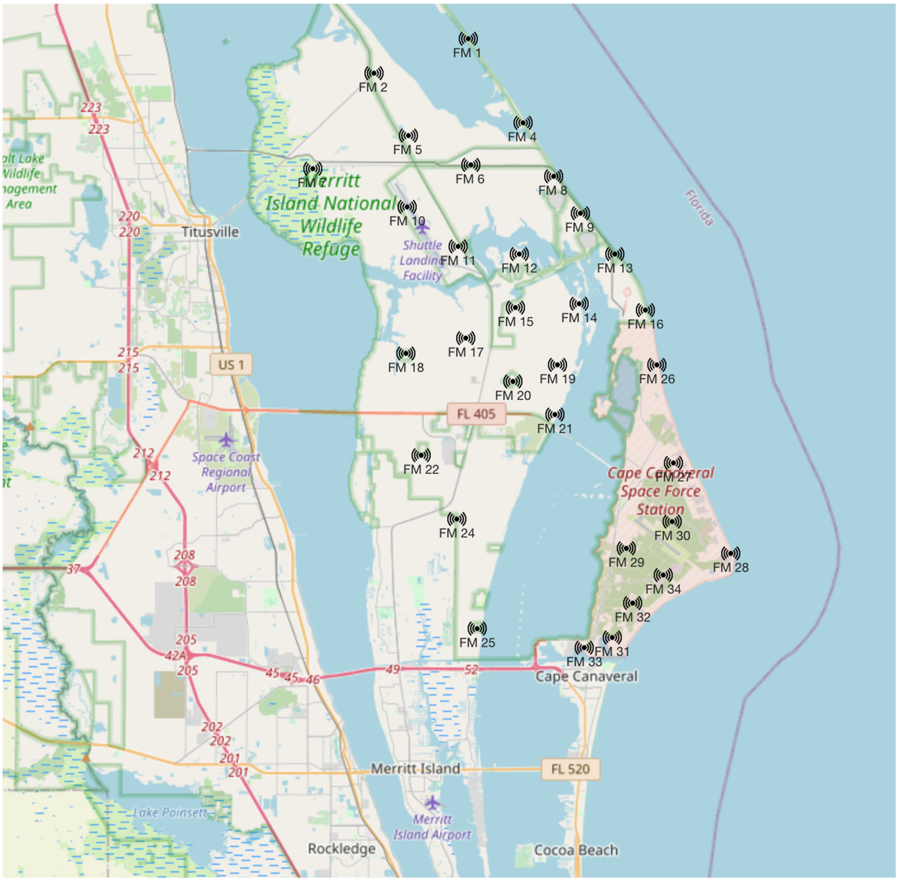
(Lucas *et al.*, 2017). The highly chaotic EFM response stored in very large datasets has confounded many attempts to create models to support lightning prediction.

Figure 3 provides a map of the KSC/CCSFS region with the location of all 31 EFM sensors. No significant shift in EFM measurements are noted at the 50 Hz rate, so the data is reduced by summarizing by the per minute mean of the 50 Hz signal to reduce overall data size.

For evaluation of the imputation method, data for field mill 25 is extracted from the main data frame. The data for field mill 25 is estimated using spatiotemporal imputation methods, and then compared against the actual observed response.

2.2 Wavelet de-noising

Wavelet techniques are used as a part of data preprocessing to reduce chaotic noise within the EFM response. Similar to the Fourier transform, a wavelet transforms model function in terms of its constituent frequencies. However, wavelet methods employ a family of unique functions that localize this approximation in time. This allows for the simultaneous approximation of a function in terms of frequency and time. Wavelet methods accomplish



Source(s): Figure courtesy of Roeder *et al.* (2015)

Figure 3.
KSC/CCSFS map with locations of EFM sensors

this by projecting approximations of a function into a series of nested subspaces, each providing a different resolution in time.

A discrete wavelet transform (DWT) can be applied to a discrete time series to produce an additive decomposition having constituent detailed time series ($\psi_{j,k}$), reflecting variations at resolution level j and a smoothed version of the time series ($\phi_{j,k}$), reflecting averages at resolution level j (Percival and Walden, 2006). Let ϕ represent the father wavelet function and ψ represent the mother wavelet function. Daubechies (1992) provides a wide variety of choices for functions which generate an orthonormal basis. With wavelets defined as

$$\phi_{j,k}(t) = 2^{j/2}\phi\left(2^{j}t - k\right) \tag{1}$$

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^{j}t - k\right) \tag{2}$$

a function of time can be represented as

$$f(t) = \sum_{j}\sum_{k} d_{j,k}\psi_{j,k}(t) + \sum_{k} s_{j_0,k}\phi_{j_0,k}(t) \tag{3}$$

where $s_{j,k} = \langle f, \phi_{j,k}\rangle$, $d_{j,k} = \langle f, \psi_{j,k}\rangle$ and $j, k \in \mathbb{Z}$. The time series is thus represented as a linear combination of the shifted and scaled versions of the wavelet functions as estimated using the wavelet coefficients $s_{j,k}$ and $d_{j,k}$. An important consequence of Equation (3) is the separation of the approximation and detailed representations of a signal.

This study employs a maximal overlap discrete wavelet transform (MODWT), a variant of wavelet transform well-suited for applications in the time series analysis. Unlike the standard DWT, which requires a dyadic sample size, the MODWT is well defined for any sized sample. Also unlike the DWT, the MODWT is shift invariant. This means that the DWT requires, and sample size in integer is a multiple of 2^j while the MODTW is defined for any sample size. Thus, the wavelet coefficients remain aligned in time with the original time series. A Haar wavelet basis is used in this implementation due to its ability to model jumps in the response signal. Figure 4 provides a visual representation of the MODWT decomposition for three detail coefficient levels and a smooth level. These properties allow the wavelet coefficients to

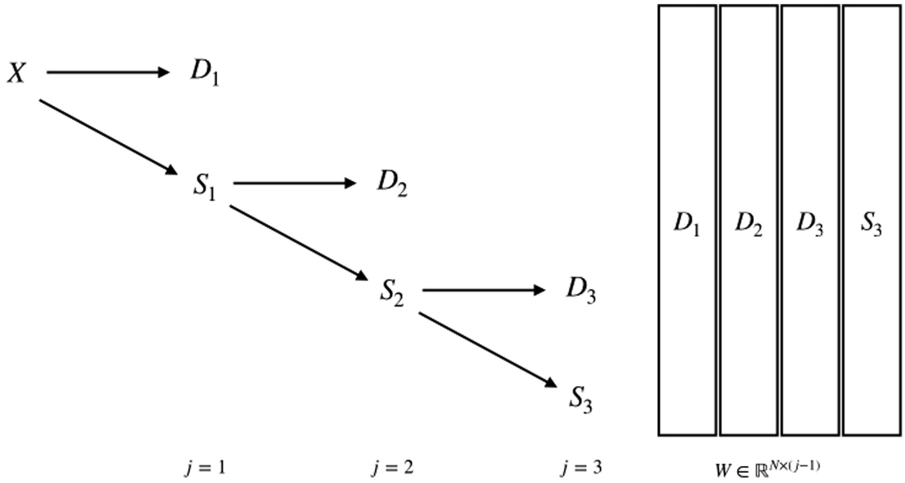


Figure 4.
Depiction of a three-level MODWT decomposition of signal X to wavelet coefficients W

Source(s): Figure courtesy of authors

remain aligned with regards to the temporal position of the original time series. However, the MODWT is a redundant transform that results in $\mathcal{O}(N\log_2 N)$ required computations or a cost of $\mathcal{O}(\log_2 N)$ when compared to the DWT.

2.2.1 Wavelet thresholding. A DWT or MODWT results in a sparse representation of the decomposed signal in the form of detail and smooth wavelet coefficient levels. This sparse approximation contains all the power of the original signal within relatively few wavelet coefficients. The remainder of the coefficients are either zero or of relatively low magnitude, and predominantly represent stochastic noise in the original time series. Thresholding manipulates these coefficients to reduce how stochastic noise is represented in the wavelet model.

This paper uses global thresholding, where a single threshold value λ is applied uniformly to all or nearly all coefficients. Consider a given threshold value λ and set

$$\widehat{f}_\lambda(t) = \sum_j \sum_k I_{\{|d_{j,k}|>\lambda\}} d_{j,k}\psi_{j,k}(t) \tag{4}$$

where I represents the indicator function (Ogden, 1997). This method is known as hard (H) thresholding, where the policy is to set coefficients to zero if less than or equal to the given value of λ. The result is that only those high magnitude coefficients are kept that represent the original signal. Then, defining the thresholded coefficients as

$$\widehat{d}_{j,k} = \delta_\lambda(d_{j,k}) \tag{5}$$

allows for representation of the hard (H) thresholding rules as

$$\delta_\lambda^H(x) = \begin{cases} x & \text{if } |x| > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Donoho and Johnstone (1994) propose an alternative method of soft (S) thresholding defined as

$$\delta_\lambda^S(x) = \begin{cases} x-\lambda & \text{if } x > \lambda \\ 0 & \text{if } |x| \leq \lambda \\ x+\lambda & \text{if } x < -\lambda \end{cases} \tag{7}$$

Soft thresholding is similar to hard methods, but values are shrunk toward zero by an amount equal to the threshold λ (Ogden, 1997).

2.3 Spatiotemporal modeling

Spatiotemporal modeling assumes a Gaussian spatiotemporal random field $\mathbb{Z}$ defined over a spatial domain $\mathcal{S}$ and a temporal domain $\mathcal{T}$ (Gräler *et al.*, 2016). A vector of samples $\mathbf{z} = (z(s_1,t_1),\ldots,(z(s_n,t_n))$ is then a collection of n measurements at distinct locations and times $(s_1,t_1),\ldots,(s_n,t_n) \in \mathcal{S}\times\mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}$ (Gräler *et al.*, 2016). Measurements may include repeated values over time for the same location or multiple values for various locations at the exact same time. Estimated values for unmeasured points (s_0, t_0) can be made since z can be assumed to be the realization of a spatiotemporal random function.

Spatiotemporal kriging is a modeling approach that produces estimated values for unmeasured locations and time using the values from the surrounding area. The method is named after Danie Krige, who developed the technique to improve the accuracy of predicting the location of underground ore reserves (Armstrong, 1998). Kriging requires the assumption that the response is a continuous random variable over the region of interest $\mathcal{S}\times\mathcal{T}$ (Robinson and Hamann, 2011). Furthermore, this modeling approach requires an assumption of stationary and spatially isotropic values across the domain of interest (Gräler *et al.*, 2016).

This means independence between the univariate probability, equal probability of occurrence regardless of location and the bivariate probability law, where the value of the underlying random function between two points depends only upon their relative distance (Isaaks and Srivastava, 1989).

The field $\mathbb{Z}$ can then be characterized with a covariance function C_{st}, where covariance depends only upon distance $h \in \mathbb{R}$ and time $u \in \mathbb{R}$ (Gräler *et al.*, 2016). The general spatiotemporal covariance function can thus be given as

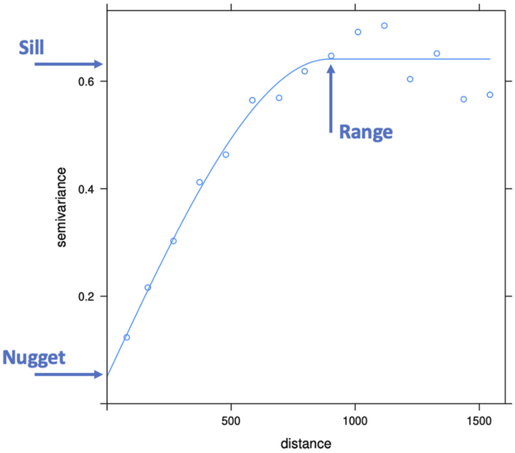
$$C_{st}(h, u) = \mathrm{Cov}\left(Z(s, t), Z(\tilde{s}, \tilde{t})\right) \tag{8}$$

for any pair of points $(s, t), (\tilde{s}, \tilde{t}) \in \mathcal{S} \times \mathcal{T}$ where $||s - \tilde{s}|| = h$ and $|t - \tilde{t}| = u$ (Gräler *et al.*, 2016).

Kriging modeling parameters retain the original nomenclature from geostatistics as seen in Figure 5. The nugget effect is the point at which the semivariogram intersects the y-axis representing semivariance. Although ideally a semivariogram would intersect at the origin, in application measurement error may result in variance amongst spatially similar measurements. The nugget effect could also be due to variations at distances smaller than the sampling distances. The range is the distance at which the semivariogram function levels off, representing the distance at which measurements are no longer autocorrelated. The sill is the value of semivariance for the range.

In practice, the covariance is modeled using a series of variograms. Model estimation is performed using the gstat package for R (Pebesma, 2004; Gräler *et al.*, 2016). First, the observed data are used to derive an empirical variogram that depicts the spatial and temporal autocorrelation of the sample points. This empirical variogram is then used as an input to a fitting routine for a generalized variogram model capable of describing covariance at varying spatial distances and times.

There are classes of generalized covariance models such as the separable covariance model, product-sum model, metric covariance model, sum-metric covariance model and simplified sum-metric covariance model (Gräler *et al.*, 2016). Each class includes a tradeoff between required assumptions and computational complexity. For instance, the separable covariance model assumes spatiotemporal covariance can be represented as



Source(s): Figure courtesy of authors

Figure 5.
Example spatial semivariogram plot from gstat package (Gräler *et al.*, 2016; Pebesma, 2004) annotated to include location of key kriging parameters nugget, sill and range

$$C_{sep}(h,u) = C_s(h)C_t(u)$$

or the product of the spatial and temporal term (Gräler *et al.*, 2016). This result in the variogram represented as

$$\gamma_{sep}(h,u) = \text{sill}\cdot\left(\overline{\gamma}_s(h) + \overline{\gamma}_t(u) - \overline{\gamma}_s(h)\overline{\gamma}_t(u)\right)$$

with standardized spatial and temporal variograms, $\overline{\gamma}_s$ and $\overline{\gamma}_t$, with separate nugget effects and joint sill of 1 (Gräler *et al.*, 2016). This study employs the Simple Sum-Metric model as it provides the best prediction values. This modeling approach assumes identical spatial and temporal covariance functions only with spatio-temporal anisotropy (Gräler *et al.*, 2016). Space and time are then matched using an anisotropy correction κ. The Simple Sum-Metric model is calculated by

$$\gamma_{ssm}(h,u) = \text{nug}\cdot\mathbf{1}_{h>0\vee u>0} + \gamma_s(h) + \gamma_t(u) + \gamma_{joint}\left(\sqrt{h^2 + (\kappa\cdot u)^2}\right)$$

which uses a single nugget effect for the spatial, temporal and joint variograms (Gräler *et al.*, 2016).

The stationary assumption of ordinary kriging further implies an assumption for an unknown and constant mean over a search neighborhood about the estimation point. This differs from simple kriging which assumes a known mean over the entire domain of interest. Ordinary kriging is a best linear unbiased estimator of an estimated point $\widehat{z}(s_0, t_0)$ as

$$\widehat{z}(s_0,t_0) = \sum_{i=1}^{n} w_i * z(s_i,t_i)$$

where w_i are the spatiotemporal kriging weights, which are allowed to change across time and location (Isaaks and Srivastava, 1989). The optimal kriging weights are then found via a search neighborhood of n points about the estimation point by solving the system of equations

$$\begin{cases} \sum_{j=1}^{n} w_j C_{st}(s_i - s_j, t_i - t_j) + \mu = C_{st}(s_i - s_0, t_i - t_0), \forall i = 1,\ldots n \\ \sum_{i=1}^{n} w_i = 1 \end{cases}$$

where μ is the Lagrange parameter (Isaaks and Srivastava, 1989; Ruybal *et al.*, 2019). Representing the ordinary kriging system of equations in matrix form results in

$$\begin{array}{ccccc} \mathbf{C} & \cdot & \mathbf{w} & = & \mathbf{D} \\ \underbrace{\begin{bmatrix} \tilde{C}_{11} & \ldots & \tilde{C}_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{C}_{n1} & \ldots & \tilde{C}_{nn} & 1 \\ 1 & \ldots & 1 & 0 \end{bmatrix}}_{(n+1)\times(n+1)} & \cdot & \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_n \\ u \end{bmatrix}}_{(n+1)\times 1} & = & \underbrace{\begin{bmatrix} \tilde{C}_{10} \\ \vdots \\ \tilde{C}_{n0} \\ 1 \end{bmatrix}}_{(n+1)\times 1} \end{array}$$

whose solution, in the form $\mathbf{w} = \mathbf{C}^{-1}\cdot\mathbf{D}$, yields the kriging weights (Isaaks and Srivastava, 1989).

3. Imputation results and discussion

This new methodology is evaluated by applying it to the EFM dataset. First, the raw EFM data are summarized to the minute to reduce the overall size of the EFM data structure. Time series data for field mill 25 is removed and stored for later comparison against the estimates produced by spatiotemporal kriging.

A MODWT transform is applied to each individual EFM time series, hard thresholding applied and an inverse MODWT is conducted to reproduce the de-noised time series. This preprocessing step reduces chaotic noise within the time series, facilitating more accurate and efficient convergence in later machine learning and artificial intelligence applications.

Spatiotemporal modeling is accomplished using the gstat package. An empirical spatiotemporal variogram is estimated from the EFM dataset. All available variogram models in the gstat package are fit and assessed. The simple sum-metric model results in the best fit by RMSE, and is thus chosen for application. Spatiotemporal kriging is then applied to interpolate values for the geodesic position of the missing field mill site.

Figure 6 provides a visual example of the estimated response (red) against the actual observed response (black). Despite the chaotic nature of EFM data, the spatiotemporal modeling technique reconstructs much of the signal for field mill 25, with an observed the mean squared error (MSE) of 0.474 and root mean squared error (RMSE) of 0.688, which in our experience with this data is a very good fit to the data. Many of the perturbations in the response are captured and modeled correctly, if not always to the full magnitude of the original observed response. This is possibly due to either the chaotic nature of the EFM data or wavelet thresholding. However, this may be a desirable property as the interpolated signal is relatively smooth and well-behaved in comparison to the chaotic raw signal. The benefit of this smoothing would depend entirely on the impact on any further application using machine learning or artificial intelligence.

Some modeling formulations using EFM for lightning forecasting employ mean imputation to fill for periods of lost sensor data. Mean imputation applies the mean of the existing time series to missing timestamped data points. Although this method appears to provide MSE of 0.6651 and RMSE of 0.8155, the constant response fails to provide any of the signal perturbations indicative of impending lightning activity. Furthermore, the relatively high assessed levels of MSE and RMSE are simply due to the EFM signal predominantly existing at a steady state measurement. The spikes out of steady state are the artifacts of interest in EFM applications, and are the indicators required in forecasting using machine learning or artificial intelligence.

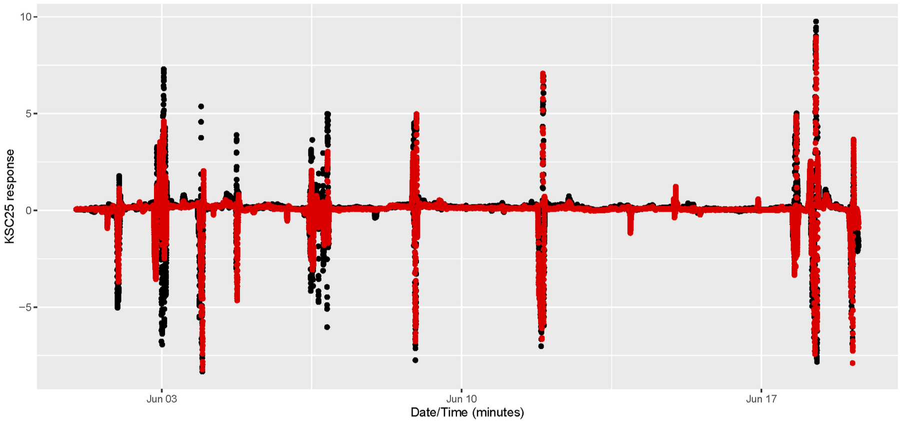


Figure 6. Observed data for field mill 25 (black) and estimated values (red) using a simple sum-metric model and spatiotemporal kriging for 1–19 June 2013, MSE = 0.474 and RMSE = 0.688

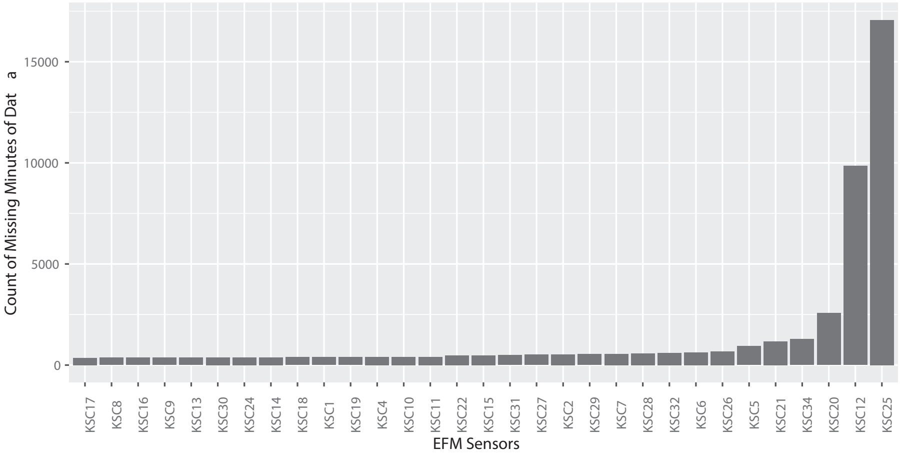
Source(s): Figure courtesy of authors

4. Application of imputed data

The fully estimated datasets are applied to the methodology of Nystrom *et al.* (2021) to evaluate the impact of using a fully imputed dataset. This methodology uses the same EFM data but with greatly reduced range of the time series to only those periods with a high proportion of EFM sensors active. Large blocks of data estimated by mean imputation caused the model to behave erratically. The application in this study seeks to apply the methodology using spatiotemporal imputation and without regard for any periods of EFM inactivity. Figure 7 provides the count of missing data points by minute for the EFM network in June 2013 as used in Nystrom *et al.* (2021). A majority of the sensors are missing data from short periods of less than 30 min when the entire network is inoperable. Linear interpolation is used to complete these time series, as there is no data available for more complex interpolation. The spatiotemporal kriging methodology is then applied to the remaining time series to interpolate missing values.

Table 1 provides the results of lightning prediction using both mean imputation and spatiotemporal imputation on the original EFM dataset. Results are presented in a confusion matrix, where the predicted state of no lightning “0” or lightning “1” is paired against actual lightning conditions observed for the same period at KSC/CCSFS. Model results predicting a lack of lightning are comparable between the two datasets. Spatiotemporal imputation results in a marked increase in the prediction accuracy for the presence of lightning (1,1) from 92.5% to 95.6%. Furthermore, this lowers the false alarm rate (1,0) that could reduce the operational impact of unnecessary lightning warnings. These improvements in model performance both increase safety for launch conditions and increase operational efficiency of launch and space flight line activities. This increase in accuracy is most likely due to the preservation of perturbations within the EFM dataset using spatiotemporal kriging, providing the semi-parametric model the key indicators for impending lightning activity.

Figure 8 provides a visual representation of the model’s prediction response against the actual observed lightning at KSC/CCSFS for 10–30 June 2013. This predicted response is estimated using the spatiotemporal imputed EFM dataset. The model provides a predictive response to nearly all the observed lightning, with three apparent false alarms during the period. Some further analysis indicates the false alarm predictions align with lightning



Note(s): Sensor KSC25 is missing the most with 17,061 min of missing data, or about 39% of all data for the month

Source(s): Figure courtesy of authors

Figure 7.
Count of missing data by minute for all 31 EFM sensors in June 2013, sorted by count

storms within the KSC/CCSFS region that did not produce lightning within the lightning warning circle under consideration for the model. Future extensions of this work will focus on reducing the impact of regional lightning storms.

Table 2 provides the results of a naïve model based upon persistence, where the model predicts the state of lightning for time $t + 1$ based exclusively on the state of lightning at time t.

Table 1. Confusion matrices for model predictions using EFM data 60 min prior to any observed lightning within the central cape lightning warning circle for 28,872 observations during 10–30 June 2013

		Observed				Observed	
		0	1			0	1
Predicted	0	27,597/27,708 99.5%	111/27,708 0.5%	Predicted	0	27,578/27,708 99.5%	130/27,708 0.5%
	1	87/1,164 7.5%	1,077/1,164 92.5%		1	51/1,164 4.4%	1,113/1,164 95.6%
Mean imputation				Spatiotemporal imputation			

Note(s): A prediction or observed value of “0” corresponds to no lightning, whereas a “1” denotes observed or predicted lightning within the lightning warning circle. Results indicate sizable improvements in the positive identification of lightning when spatiotemporal imputation is used to complete the EFM dataset
Source(s): Table courtesy of authors

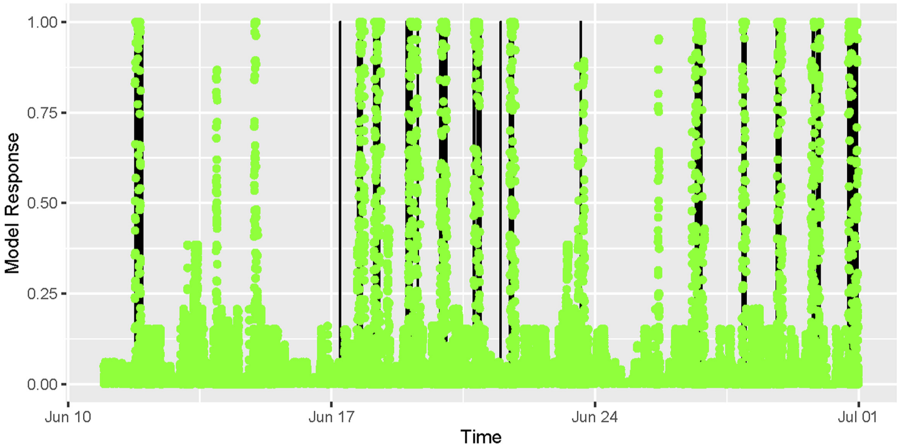


Figure 8. Predicted model response (green) using imputed EMF dataset against actual observed lightning (black) on Cape Canaveral, June 2013

Source(s): Figure courtesy of authors

Table 2. Confusion matrix for performance of the naïve persistence model

		Observed	
		0	1
Predicted	0	13,698/13,814 99.16%	116/13,814 0.84%
	1	116/586 19.8%	470/586 80.2%

Note(s): This model develops a forecast using only the lightning state of the previous timestamp. For instance, if there is no lightning at time t, then the model predicts no lightning at $t + 1$. The wavelet enabled semi-parametric modeling approach outperforms a naïve model in this implementation and indicates this new methodology has explanatory power in the prediction of lightning phenomena
Source(s): Table courtesy of authors

This manner of comparison is common in the meteorological literature, and shows whether the model under evaluation is providing explanatory insights to weather phenomena. The wavelet-enabled semi-parametric modeling approach outperforms the persistence model, most notably in the identification of the presence of lightning.

5. Conclusion

Spatiotemporal kriging provides an excellent method to recreate a missing time series that includes spatial autocorrelation. The technique proved robust, despite the chaotic nature of EFM measures of atmospheric electrostatic potential. Furthermore, the interpolated time series displays evidence of some smoothing while also preserving the signal of interest for lightning prediction. Both of these qualities may aid in convergence in additional machine learning or artificial intelligence applications while still facilitating accurate and timely predictions.

References

Aranguren, D., Inampués, J., Torres, H., López, J. and Pérez, E. (2012), "Operational analysis of electric field mills as lightning warning systems in Colombia", *2012 International Conference on Lightning Protection (ICLP)*, IEEE, pp. 1-6.

Armstrong, M. (1998), *Basic Linear Geostatistics*, Springer-Verlag Berlin Heidelberg, New York, NY.

Cheng, A. (2020), "Lightning prediction for space launch using machine learning based off of electric field mills and lightning detection and ranging data", Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics (SIAM), Vol. 61.

Donoho, D.L. and Johnstone, J.M. (1994), "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, Vol. 81 No. 3, pp. 425-455.

Eshel, G. (2012), *Spatiotemporal Data Analysis*, Princeton University Press, Princeton, NJ.

Gräler, B., Pebesma, E. and Heuvelink, G. (2016), "Spatio-Temporal Interpolation using gstat", *The R Journal*, Vol. 1 No. 8, pp. 204-218.

Hill, D.E. (2018), "Lightning prediction using artificial neural networks and electric field mill data", Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Isaaks, E.H. and Srivastava, R.M. (1989), *An Introduction to Applied Geostatistics*, Oxford University Press, New York, NY.

Junger, W. and De Leon, A.P. (2015), "Imputation of missing data in time series for air pollutants", *Atmospheric Environment*, Vol. 102, pp. 96-104.

Krider, E.P. (1989), "Electric field changes and cloud electrical structure", *Journal of Geophysical Research: Atmospheres*, Vol. 94 No. D11, pp. 13145-13149.

Lopez, J., Perez, E., Herrera, J., Aranguren, D. and Porras, L. (2012), "Thunderstorm warning alarms methodology using electric field mills and lightning location networks in mountainous regions", *2012 International Conference on Lightning Protection (ICLP)*, IEEE, pp. 1-6.

Lucas, G.M., Thayer, J.P. and Deierling, W. (2017), "Statistical analysis of spatial and temporal variations in atmospheric electric fields from a regional array of field mills", *Journal of Geophysical Research: Atmospheres*, Vol. 122 No. 2, pp. 1158-1174.

Moritz, S. and Bartz-Beielstein, T. (2017), "imputeTS: time series missing value imputation in R", *The R Journal*, Vol. 9 No. 1, p. 207.

Nystrom, J.K. (2021), "Wavelet methods for very-short term forecasting of functional time series", PhD thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Nystrom, J., Hill, R.R., Geyer, A., Pignatiello, J. J., Jr. and Chicken, E. (2021), "Experimental Design in Complex Model Formulation for Lightning Prediction", *International Journal of Experimental Design and Process Optimisation*, Vol. 6 No. 4, pp. 304-332.

Ogden, R.T. (1997), *Essential Wavelets for Statistical Applications and Data Analysis*, Springer Science & Business Media, New York, NY.

Pebesma, E.J. (2004), "Multivariable geostatistics in S: the gstat package", *Computers and Geosciences*, Vol. 30 No. 7, pp. 683-691.

Percival, D.B. and Walden, A.T. (2006), *Wavelet Methods for Time Series Analysis*, Vol. 4, Cambridge University Press, New York, NY.

Robinson, A.P. and Hamann, J.D. (2011), *Forest Analytics with R: An Introduction*, Springer Science & Business Media, New York, NY.

Ruybal, C.J., Hogue, T.S. and McCray, J.E. (2019), "Evaluation of groundwater levels in the arapahoe aquifer using spatiotemporal regression kriging", *Water Resources Research*, Vol. 55 No. 4, pp. 2820-2837.

Speranza, D. (2019), "Lightning prediction using recurrent neural networks", Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Further reading

Roeder, W.P., Cummins, B.H., Cummins, K.L., Holle, R.L. and Ashley, W.S. (2015), "Lightning fatality risk map of the contiguous United States", *Natural Hazards*, Vol. 79 No. 3, pp. 1681-1692.

Corresponding author

Raymond R. Hill can be contacted at: rayrhill@gmail.com