An Analysis of Imputation for Weather Data

June 17, 2025

Abstract

Predicting future weather in a specific location requires precise historical weather data from that area. A key challenge in this domain is dealing with incomplete data (i.e. missing values). This study will examine how the amount of missing data and the imputation methods used influence a model's prediction. Kalman Smoother, moving average, and linear interpolation are tested against each other at 10%, 25%, and 40% missing data using Monte Carlo simulation to generate the data and test an long-sthort-term memory (LSTM) model. By comparing the root mean squared error (RMSE) statistics between each group, we are able to find the best method for imputing time-series data.

Introduction

Weather forecasting models are widely used by the general public, usually to help us make decisions about what to wear or to help plan future activities. However, these models have a much larger impact than we realize. They can often influence agricultural practices, public policy, and resource management. Because of this, it is important that we deploy models that can accurately represent how the weather evolves. The challenge is that weather systems are complex and exhibit chaotic behavior, meaning that small changes in initial conditions result in large differences over time. This makes implementing such models difficult.

Another key challenge is dealing with incomplete (missing) data. When working with time-series data, it is critical that we have access to the full data. However, it is rarely the case that we have access to the complete observations. Weather stations can randomly break down or have routine maintenance scheduled. In such cases, our only option is to turn to imputation methods to fill the data. The goal of this experiment is to examine how the predictions of a model are influenced by the following factors:

- The amount of missing data
- The imputation methods used to fill the data

with a primary focus on the latter, since that is the one we can control.

Data & Model

To perform this experiment, we will use data from NOAA's Climate Data Online Search Tool [1]. We will specifically focus on data collected by Denver Centennial Airport (APA) station [2] starting on January 1, 2005 and ending on April 3, 2025. The data consists of many variables (temperature, precipitation, wind, etc.), but we will only consider the maximum temperature column (`TMAX`) for our time-series analysis. We impute the few missing values (14) in `TMAX` using the previous value and leave the more interesting imputation methods for the experiments later on. The data is then split into training (70%), validation (15%), and testing (15%) sets. Finally, the data is normalized using `MinMaxScaler()` from `scikit-learn` to map our values between 0 and 1.

We will make predictions using a specialized recurrent neural network (RNN) known as long-short-term memory (LSTM)[3]. We construct this model in `PyTorch`[4] to have one layer with 16 hidden units. This model will take the previous week's maximum temperatures as input and output the next day's maximum temperature. The model is trained over 30 epochs using the mean square error (MSE) as its loss function and a learning rate of 10^{-3}. Using this model, we can get a decent estimate of the tomorrow's temperature.

Methods

In this section, we will discuss the Monte Carlo simulation. We begin by generating the data. To simulate extended maintenance outages, we will insert sequences of five consecutive missing data points into the training and validation data only. The data sets will have roughly 10%, 25%, and 40% of their values missing. For the training set, this corresponds to 104, 259, and 414 sequences respectively. For the validation set, this corresponds to 22, 56, and 89 sequences respectively. We will impute the data using the following techniques: Kalman Smoother (KS)[5, 6], 3 day moving average (MA)[7], and linear interpolation (LI)[8].

We denote the number of data points in the data set as N (indexed from 0 to $N-1$). To generate the data, we sample indices from the set $\{3, 8, 13, \ldots, 5(n-1)+3\}$ without replacement where n is the quotient

of $N-2$ and 5. This is critical since the imputation methods used will rely on having the first three data points or the final data point available. Then at each sampled index i, we insert missing values at indices $\{i, i+1, i+2, i+3, i+4\}$. We create three new data sets by imputing using the three methods above. This is repeated 15 times for each imputation method and level of missing data on the training and validation datasets (135 training and validation sets each).

For each training and validation pair, we initialize a model with the same architecture from earlier, and train it using using the same training regime from the Data & Model section. The model is tested on the testing data (which is not altered). To evaluate our performance on the true test data, we will use the root mean squared error (RMSE). To control for variability, we initialize each model with the same weights.

Results & Discussion

We show the RMSE statistics in Table 1 and visualize it using a boxplot in Figure 1.

Imputation Method	**KS**			**MA**			**LI**		
% Missing Data	**10%**	**25%**	**40%**	**10%**	**25%**	**40%**	**10%**	**25%**	**40%**
RMSE μ	9.313	9.352	9.363	9.318	9.346	9.383	9.302	9.316	9.358
RMSE σ	0.011	0.040	0.033	0.010	0.028	0.030	0.007	0.015	0.014

Table 1: RMSE mean and standard deviations are shown for each of the nine dataset configurations. Each dataset configuration was generated 15 times and used to train a model.

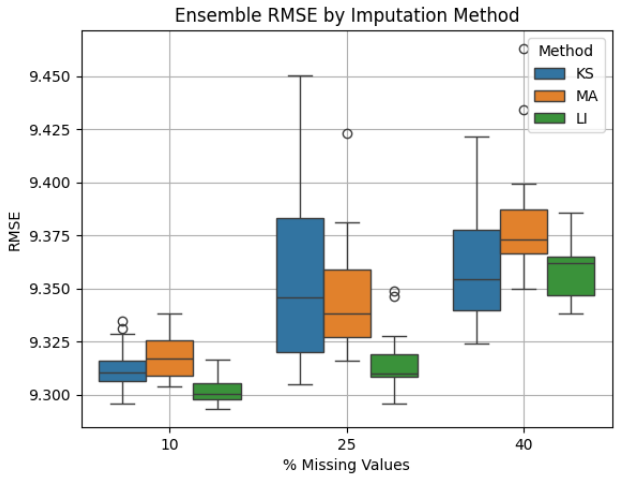

Figure 1: Box plot of RMSE for each imputation method and level of missing values. LI has the lowest median for 10% and 25% missing data while KS has the lowest RMSE median for 40% missing data.

To test for differences between the imputation methods, we perform two way analysis of variance (ANOVA). The results are shown in Table 2. For $\alpha = 0.05$, we would reject the null hypothesis and conclude that there is a statistically significant effect of imputation method and amount of missing data which makes sense intuitively as these two factors heavily influence the data. More interestingly, we find that there is not a statistically significant interaction between the two.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value	η^2
Imputation Method	0.013	2	0.007	12.018	0.000	0.160
% Missing	0.074	2	0.037	65.718	0.000	0.511
Imputation Method × % Missing	0.005	4	0.001	2.174	0.076	0.065
Residual	0.071	126	0.001	—	—	—

Table 2: Two way ANOVA results

% Missing Data	Imputation 1	Imputation 2	Standard Error	T-stat	p-value
10%	KS	LI	0.003	3.210	0.011
	KS	MA	0.004	-1.189	0.469
	LI	MA	0.003	-4.869	0.000
25%	KS	LI	0.011	3.310	0.010
	KS	MA	0.013	0.472	0.885
	LI	MA	0.008	-3.671	0.003
40%	KS	LI	0.009	0.604	0.820
	KS	MA	0.011	-1.679	0.231
	LI	MA	0.009	-2.896	0.023

Table 3: Games-Howell post hoc pairwise test results.

Next, we performed the Games-Howell test to compare the imputation methods pairwise. These results are shown in Table 3. Correcting α for the number of tests gives $\alpha = \frac{0.05}{9} \approx 0.006$. Using this significance level, we find statistical differences for LI and MA in the 10% and 25% groups.

At all levels of missing data, LI has the lowest RMSE mean. This is quite interesting since LI is the most simple method of the three: s straight line between two points. In this case, where the sequence length is not very long, the principle of Occam's razor holds, and LI is enough to capture the trend in the data. However, our sample size is quite small, so these statistics may not accurately represent the larger picture. Furthermore, the results may not hold if the sequences span long periods where trends are no longer linear.

Nevertheless, given the available data, it appears that LI is enough—even preferred—for imputing short sequences of missing data. However, to draw more conclusive insights, we must address the limitations of this study design. The most glaring issue with this study is the small sample size. Unfortunately, our study was limited by computational resources, making larger sample sizes difficult to manage. This study is also limited to a single geographic location. In areas with relatively stable temperature patterns, LI would reasonably do a good job of imputing missing data. Whether or not this holds in areas that exhibit more chaotic temperature patterns remains to be seen. Expanding this study to multiple geographic locations with larger sample sizes would greatly increase the reliability of our results. Another issue of this study is that it uses only one type of model. It is possible that the imputation method is more dependent on the model than the data. This study could be extended to include multi-variate and multi-step time-series forecasting.

Despite the limitations of the study, we were able to find interesting results which can be explored in future studies. This study establishes the framework for future studies to build upon as we tackle the issue of imputation in time-series data.

References

[1] National Centers For Environmental Information (NCEI), "Global historical climatology network - daily (ghcnd)," https://www.ncei.noaa.gov/cdo-web/search?datasetid=GHCND, 2025, accessed: 2025-06-17.

[2] National Centers for Environmental Information (NCEI) / NOAA, "Station details for ghcnd:usw00093067," https://www.ncei.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00093067/detail, 2025, accessed: 2025-06-17.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[4] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[5] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, p. 35–45, Mar. 1960. [Online]. Available: http://dx.doi.org/10.1115/1.3662552

[6] H. E. RAUCH, F. TUNG, and C. T. STRIEBEL, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, p. 1445–1450, Aug. 1965. [Online]. Available: http://dx.doi.org/10.2514/3.3166

[7] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2008, available at https://otexts.com/fpp2/moving-averages.html.

[8] "Linear Interpolation Formula - Derivation, Formulas, Examples — cuemath.com," https://www.cuemath.com/linear-interpolation-formula/, accessed: 17-06-2025.