Predicting Meteorological Values on a Spatial Grid

Felipe Hernández
University of Pittsburgh
felher.c@gmail.com

Ben Humberston
Carnegie Mellon University
bhumbers@cs.cmu.edu

1 Introduction

1.1 Motivation

Weather forecasts are broadly used; from personal activity planning (What clothes should I wear today? Will it be a good time to plan outdoor events?); to large-scale economic decision-making (What activities should be prioritized on a crop next week? What precautions should the air-traffic controllers enforce on a given day?); to emergency preparation and response (What alternate routes should become available due to snow?, Where should the emergency vehicles be sent during a flood event?). The availability and accuracy of forecasts thus have a profound impact on human activities at many levels, both in measurable and unmeasurable aspects. However, predicting weather is a difficult research problem. Most often, physically-based models with global scale are used to forecast future conditions. In this project, we will instead take a machine learning approach focused on a local scale and attempt to predict atmospheric variables at specific geographic locations. The forecasts will be based on prior atmospheric system state in the neighborhood of a selected location. In particular, we will attempt to predict system variables such as pressure, precipitation, and temperature based on the previous values of these variables provided by regular historical satellite observations.

1.2 Related Work

Atmospheric sciences have always studied a variety of methods for weather forecasting. Many physical phenomena are involved in the state of the weather variables at a certain time, including energy and mass transfer between the sun, the different layers of the atmosphere, the ground, the oceans, etc. The development of physically-based forecasting models requires the incorporation of a wide variety of phenomena, such as used in [7] and [4]. The availability of meteorological measurements has helped the development of these methods, as well as other data-driven approaches. Machine Learning approaches have been used with isolated gauge data ([5] and [6]); more recently, it has also been applied to the data from emerging radar [3] and satellite [10] technologies.

One such satellite-based data product is the NLDAS (North American Land Data Assimilation System), a service hosted by the Goddard Earth Sciences Data and Information Services Center at NASA [2]. It provides hourly weather data for the US beginning in 1980. The data is provided on a regular grid with a resolution of 1/8th of a degree in latitude-longitude coordinates. Table 1 shows the list of variables available from this data source.

2 Method

2.1 Linear Regression

Linear regression was used as an initial naïve approach to estimate the values of the weather variables. For each cell in the NLDAS-2 map, we have a value y to predict, and a set of variables $\boldsymbol{x}$ which are inputs into the prediction. We train a linear regression model of the form $y = \boldsymbol{w}^T \boldsymbol{x}$, where y is a linear combination of the values in vector $\boldsymbol{x}$. The vector $\boldsymbol{w}$ contains the coefficients of the linear model.

Name	Description	Units
precipitation	Accumulated height of precipitated water column	millimeters
available potential energy	180-0 mb above ground Convective Available Potential Energy	Joules per kilogram
% convective precipitation	Fraction of total precipitation that is convective	none
LW radiation	Long wave radiation flux downwards (surface)	watts per square meter
SW radiation	Short wave radiation flux downwards (surface)	watts per square meter
potential evaporation	Potential evaporation	millimeters
surface pressure	Surface pressure	pascals
specific humidity	2 m above ground Specific humidity	none
temperature 2 m	2 m above ground Temperature	kelvin
zonal wind speed	10 m above ground Zonal wind speed	meters per second
meridional wind speed	10 m above ground Meridional wind speed	meters per second

Table 1: Meteorological variables available from NLDAS-2

If we take a time frame made of T one-hour steps, and look at all the n cells in the study area, we have $T * n$ examples of y values as a function of the corresponding values of $\boldsymbol{x}$. These values can be arranged in a vector $\boldsymbol{y}$ of size $T * n$, and a matrix $\boldsymbol{X}$ of size $(T * n) \times f$, with f being the number of features in the linear model. We can compute the coefficient vector $\boldsymbol{w}$ using the following equation:

$$\boldsymbol{w} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{1}$$

Additionally, using the same formulation, a polynomial regression can be computed by adding higher degree terms into the set of features.

2.2 VAR model

The linear regression model is intended as a baseline solution. Our main work uses a vector autoregressive (VAR) model in order to simulate and predict the evolution of the meteorological system over time. We adopt the formulation and estimation strategy given in [1]. The model describes the system values at time t given some number of prior system states p, which is the lag order of the model:

$$\boldsymbol{X}_t = \boldsymbol{c} + \boldsymbol{\Pi}_1 \boldsymbol{X}_{t-1} + ... + \boldsymbol{\Pi}_p \boldsymbol{X}_{t-p} + \epsilon_t, \tag{2}$$

Here, $\boldsymbol{X}_i$ is the system state at the i^{th} timestep; each of these is an N-length vector formed by reshaping and concatenating the various geographic grids of observed data variables (eg: temperatures, pressures, and so on at each of the grid points). Note that the data is offset to zero-mean over all time steps and normalized to have unit variance. $\boldsymbol{c}$ is a constant offset. Each $\boldsymbol{\Pi}_j$ specifies the $N \times N$ model coefficient matrix which defines weighted sums of the observed data at timestep $t - j$. In a model of lag order p, the new state $\boldsymbol{\Pi}_t$ is a sum of linear combinations of prior system states $\boldsymbol{\Pi}_j \boldsymbol{X}_{t-j}$ from timestep $t - 1$ to $t - p$. Finally, ϵ_t is a zero-mean noise generator which is uncorrelated between time steps which introduces innovations to the process.

This system can be solved for the unknown $\boldsymbol{\Pi}_j$ matrices using ordinary least squares by rewriting it as a system of equations where each row is the equation for one system variable at one time step:

$$\boldsymbol{x}_i = \boldsymbol{Z} \boldsymbol{\pi}_i + \boldsymbol{e}_i, i = 1, ..., N \tag{3}$$

Given that we are training on T timesteps, $\boldsymbol{x}_i$ is a T-length vector giving the values for the i^{th} system variable over all timesteps (eg: "atmospheric pressure at (20° N, 34° W) over a 24 hour period"). $\boldsymbol{Z}$ is a $T \times k$ matrix ($k = np + 1$), where the t^{th} row is defined as $\boldsymbol{Z}_t = (1, \boldsymbol{X}_{t-1}^T, \boldsymbol{X}_{t-2}^T, ..., \boldsymbol{X}_{t-p}^T)$, $\boldsymbol{e}_i$ is a zero-mean noise vector, and $\boldsymbol{\pi}_i$ is a k-length vector of coefficients to be determined. Using a typical least squares approach, we create an estimate $\hat{\boldsymbol{\pi}}_i$ for $i = 1, ..., N$, yielding our estimation for the full model:

$$\hat{\boldsymbol{\Pi}} \in (k \times N) = \begin{bmatrix} \hat{c} \\ \hat{\boldsymbol{\Pi}}_i^T \\ \vdots \\ \hat{\boldsymbol{\Pi}}_p^T \end{bmatrix} = [\hat{\boldsymbol{\pi}}_1 \quad \dots \quad \hat{\boldsymbol{\pi}}_N] \tag{4}$$

2.3 Lasso

Our data set includes a number of metereological variable types which are defined on the geographic grid. Since we suspect that not every data observation will be critical to modeling the system's evolution, we intend to use the lasso regularization [8] to enforce sparse feature selection. This changes our least squares problem slightly. We now wish to minimize the quantity

$$\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{Z}\boldsymbol{\pi}_i)^2 + \lambda^{(1)} \sum_{i=1}^{k} \sum_{j=1}^{N} |\boldsymbol{\Pi}_{i,j}| \tag{5}$$

$\lambda^{(1)}$ is the Lagrange multiplier on the lasso constraints. This is a convex optimization and should be solvable using standard tools.

2.4 Fused Lasso

Additionally, we have reason to believe that the model coefficients $\boldsymbol{\Pi}$ should exhibit some temporal consistency so that the relationship of parameters within ranges of timesteps $\boldsymbol{\Pi}_{t1}...\boldsymbol{\Pi}_{t2}$ show constant trends. To explore this, we turn to the fused lasso, a modified version of lasso introduced by [9]. It is useful for time series data because it encodes the relationship between successive parameters. Our modified objective function for fused lasso is

$$\sum_{i=1}^{N}(x_i - Z\boldsymbol{\pi}_i)^2 + \lambda^{(1)} \sum_{j=1}^{k} \sum_{i=1}^{N} |\boldsymbol{\Pi}_{i,j}| + \lambda^{(2)} \sum_{i=1}^{N} \sum_{t=2}^{T} |\boldsymbol{\pi}_{i,t} - \boldsymbol{\pi}_{i,t-1}| \tag{6}$$

Here, $\lambda^{(2)}$ is the Lagrange multiplier on the additional constraints for fused lasso. Refer to Equation 4 for the relationship between $\boldsymbol{\pi}$ terms and $\boldsymbol{\Pi}$.

The fused lasso regularization may be particularly useful if we examine the VAR model behavior over longer time scales than hour-by-hour weather. There are, naturally, seasonal variations in meteorological variables, and it is possible that the system behavior encoded by the model coefficients $\boldsymbol{\Pi}$ is also dynamic based on the time of year.

2.5 Evaluation

The success of our model depends on its ability to correctly forecast (predict) the new system state at time t based on the system state at times $t - p$ to $t - 1$. Since there is some inherent noise ϵ in the variables of the system, the primary goal is to minimize the residual of predicted system state $\hat{\boldsymbol{X}}_t$ relative to the actual historical observation $\boldsymbol{X}_t$ to within some reasonable range of this noise term. The size of our data set allows us to easily train on a subset of the full data and test on another independent set (eg: train on data from the year 2004, test on data from the year 2007).

We intend to use cross-validation in order to ensure that our data does not overfit to any particular data set. Additionally the lag order p of the VAR model provides a way of controlling the model complexity; larger values of p yield more complex models. We can observe the training and test error for models trained for different lags and thus estimate an optimal lag size which is large enough to capture the system dynamics but small enough to avoid overfitting.

3 Results

Currently, we have implemented both a simple linear regression model as well as a standard VAR(p) model which predicts new system values given prior system states.

3.1 Linear Regression

A small region around Pittsburgh was selected as the test case for this method. It spans a 6 x 6 grid of 10 km x 10 km size cells. Hourly records for each cell were used from 9/15/2010 to 10/15/2010.

Four linear regression models were computed for each one of the meteorological values in NLDAS-2. The first two were determined assuming a static framework: computing the current value if all the other current values are known. The first regression only used terms of degree one, while the second used terms of degree two (the quadratic term of each value plus the all the product terms between any two values). The third and fourth models were also of degree one and two, but were developed to estimate the values of the weather variables in the next time step (an hour into the future) given the current values of all the variables in the current time step.

In order to test the performance of each model, five runs were executed randomly taking 10% of the examples as a testing set, and then computing the root mean square error (RMSE) of the predictions for both the training and the testing set. Table 2 shows the relative errors of the trained models on the test set for each variable, as the ratio between the RMSE and the standard deviation of the entire sample of each variable.

Variable	**Current time step**		**Next time step**	
	degree 1	**degree 2**	**degree 1**	**degree 2**
precipitation	101.28%	91.45%	87.33%	84.21%
potential energy	79.15%	76.41%	22.20%	18.83%
% convective	89.82%	84.78%	72.68%	73.00%
LW radiation	61.11%	59.93%	28.88%	29.39%
SW radiation	41.81%	42.32%	25.01%	24.60%
potential evaporation	32.58%	30.65%	29.34%	27.29%
surface pressure	93.67%	88.34%	8.60%	8.52%
specific humidity	47.63%	44.35%	38.11%	35.48%
temperature	46.22%	46.34%	12.61%	11.82%
wind speed	77.16%	78.18%	21.14%	19.90%
Average	**67.04%**	**64.27%**	**34.59%**	**33.30%**

Table 2: Relative testing errors of the linear regression models

As can be seen, the high relative errors for most cases indicate that the proposed approach is still not adequate for forecasting. However, there is a strong indication that variables are highly dependent on their previous values (since the estimation for the next time step is mostly better), and that there is a small improvement when higher order terms are included into the model. Additionally, there is a wide variation in the quality of the prediction between variables. There is also virtually no difference between training and testing errors, so we can conclude that the models are not over-fitted.

3.2 VAR Model

We illustrate an example of the output from the VAR model in Figure 1. Thus far, this model has also shown poor performance, where the errors in the model output are very large relative to the expected noise in each observed variable.

There are several possible reasons for this which we intend to investigate. First, the VAR model assumes a *stationary* process, and our current preprocessing of the data before fitting the model may not guarantee this assumption. Additionally, not all variables which may affect predictions are yet included in this model (eg: wind speeds, elevation, etc.). We will observe whether adding these terms improves predictive accuracy before testing the (fused) lasso-regularized version of this model.

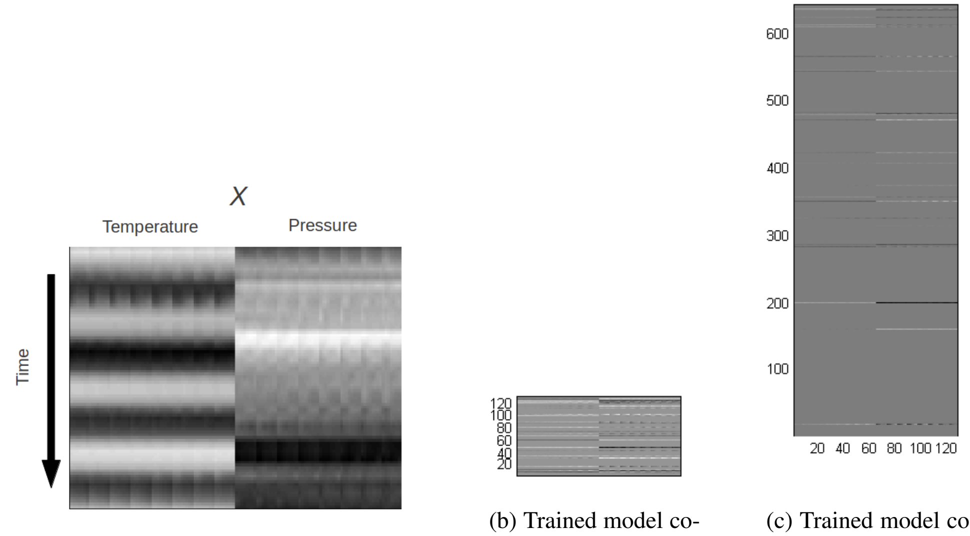
4 Appendix

4.1 Updated Timeline

- Training and performance evaluation of VAR model Week 1 (Nov. 18-24)
- Training and performance evaluation of fused lasso VAR regression Week 1 (Nov. 18-24)
- Comparison of linear regression, static VAR approach, and fused lasso regression Week 2 (Nov. 25 Dec. 1)
- Poster preparation Week 2 (Nov. 25 Dec. 1)
- Report preparation Week 3 (Dec. 2-8)

References

[1] Vector autoregressive models for multivariate time series. In *Modeling Financial Time Series with S-PLUS*, pages 385–429. Springer New York, 2006.



(a) Input data comprised of temperature and pressure values on a small grid region

(b) Trained model coefficients assuming a lag order $p = 1$

(c) Trained model coefficients assuming a lag order $p = 5$

Figure 1: An illustrative example of the input data $\boldsymbol{X}$ and trained output model $\hat{\boldsymbol{\Pi}}$ obtained using different lag parameters. The data is comprised of hourly samples over a 4-day window. Each row in $\boldsymbol{X}$ corresponds to all the observed values at a particular timestep, where the 2D gridded values for each variable type (in this case, surface temperature and pressure) are reshaped into row vectors. Note that the large, monochrome gray regions in each $\hat{\boldsymbol{\Pi}}$ indicates significant sparsity of the model, which was not expected until after we apply lasso regularization.

[2] National Aeronautics and Space Administration. Land data assimilation systems.

[3] Ralph R Ferraro and Gerard F Marks. The development of ssm/i rain-rate retrieval algorithms using ground-based radar measurements. *Journal of Atmospheric and Oceanic Technology*, 12(4):755–770, 1995.

[4] Daniel Harris, Efi Foufoula-Georgiou, Kelvin K Droegemeier, and Jason J Levit. Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology*, 2(4):406–418, 2001.

[5] Wei-Chiang Hong. Rainfall forecasting by technological machine learning models. *Applied Mathematics and Computation*, 200(1):41–57, 2008.

[6] Paras and Sanjay Mathur. A simple weather forecasting model using mathematical regression. *Indian Research Journal of Extension Education*, 1(1):161–168, 2012.

[7] W.J. Shuttleworth. *Terrestrial Hydrometeorology*. Wiley, 2012.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[9] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[10] Yubin Yang, Hui Lin, Zhongyang Guo, and Jixi Jiang. A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis. *Computers & geosciences*, 33(1):20–30, 2007.