|  |
| --- |
|  |
| **Predicting Meteorological Values on a Spatial Grid using the North American Land Data Assimilation System** |
|  |

**Felipe HernándezBen Humberston**

University of Pittsburgh Carnegie Mellon University

*felher.c@gmail.com* *bhumbers@cs.cmu.edu*

|  |
| --- |
|  |

**Abstract**

In this project we study the use of a Vector Auto-Regression (VAR) model for weather forecasting. Using satellite-based hourly meteorological observations, we explore the inclusion of spatial relationships with neighboring areas, previous time windows of different size, and weather variable interdependencies into the regression model to determine the effect of these variants on the forecasting performance. Additionally, a lasso regularization version of VAR is also used to compensate the overfitting of the model when the number of features in the model gets too large. The effects of the different model setups, as well as the overall performance of the VAR approach are shown and discussed.

**1 Introduction**

* 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.

**1.1 Background and 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. ([1]) The development of physically-based forecasting models requires the incorporation of such wide variety of phenomena and, because of this, also requires large amounts of measurements that are generally not available. Together with the large complexity of such models, the use of simplified models is often preferred or even mandatory [2].

Machine Learning approaches have been explored to construct such less complex models based on meteorological measurements. Many of these techniques have been tailored to fit the nature of the observations available. Weather monitoring stations are the most predominant source, providing point measurements with high accuracy and varying temporal resolution.

Artificial Neural Networks have proven to be effective in such cases for forecasting rainfall amounts in the near future given a time series of previously observed values ([3] and [4]). More recent works have attempted to combine ANNs with other techniques to improve the performance of predictions. In [2], a Support Vector Regression, together with a Chaotic Particle Swarm Optimizer, is used to determine the best parameters for a Recursive ANN to predict precipitation values during monsoons, based on the precipitation time series at gauging points.

More recently, satellite and Doppler radar based observations have become available, providing ubiquitous coverage but with decreased precision.

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 [9]. 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.

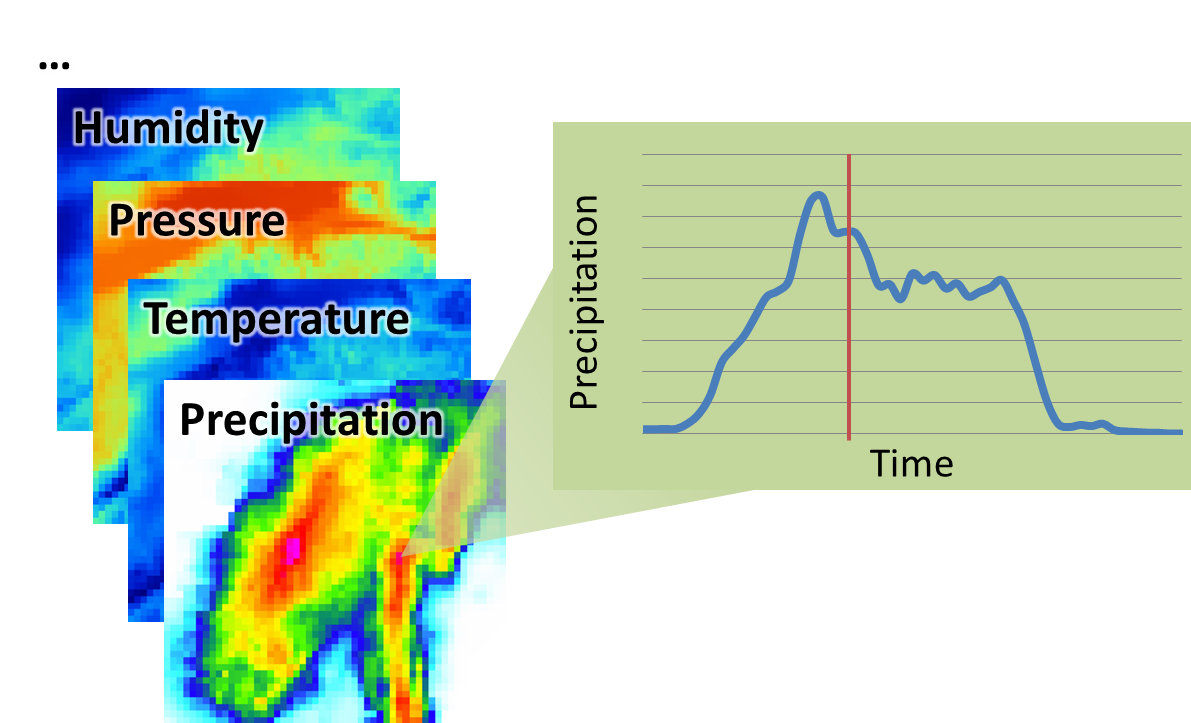


Figure 1. Spatial and temporal dimensions of the NLDAS-2 meteorological observations [9]: hourly images of each variable are available on a regular latitude-longitude grid.

Table 1 shows the list of variables available from this data source.

Table 1: Meteorological variables available on NLDAS-2 [9]

|  |  |  |
| --- | --- | --- |
| **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 |

1. **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 x which are inputs into the prediction. We train a linear regression model of the form y = wTx, where y is a linear combination of the values in vector x. The vector w contains the coefficients of the linear model.

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 *x*. These values can be arranged in a vector *y* of size *T \* n*, and a matrix *X* of size (*T \* n*) × *f*, with *f* being the number of features in the linear model. We can compute the coefficient vector *w* using the following equation:

|  |  |
| --- | --- |
|  | (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:

|  |  |
| --- | --- |
|  | (2) |

Here, ***X****i* is the system state at the *ith* time step; 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. *c* is a constant offset. Each **Π***j* specifies the *N × N* model coefficient matrix which defines weighted sums of the observed data at time step *t - j*. In a model of lag order p, the new state **Π***t* is a sum of linear combinations of prior system states **Π***j***X***t-j* from time step *t - 1* to *t - p*. Finally, *εt* 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 **Π***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:

|  |  |
| --- | --- |
|  | (3) |

Given that we are training on *T* time steps, *xi* is a *T*-length vector giving the values for the *ith* system variable over all time steps (eg: “atmospheric pressure at (20° N, 34° W) over a 24 hour period”). *Z* is a *T × k* matrix (*k = np + 1*), where the *tth* row is defined as *Zt =* (1, ***X***T*t -*1, ***X***T*t-*2, … , ***X***T*t-p*), *ei* is a zero-mean noise vector, and *πi* is a *k*-length vector of coefficients to be determined. Using a typical least squares approach, we create an estimate *πi* for *i* = 1, … , *N*, yielding our estimation for the full model:

|  |  |
| --- | --- |
|  | (4) |

We consider the effects of three factors on model performance: whether or not values of neighboring cells are included as input data (radius *r*); whether variables are estimated independently or whether all variables are used as interdependent inputs to the model; and the number of previous states included in the model (VAR lag order *p*). Figure 2 shows the different dependency models to be explored.

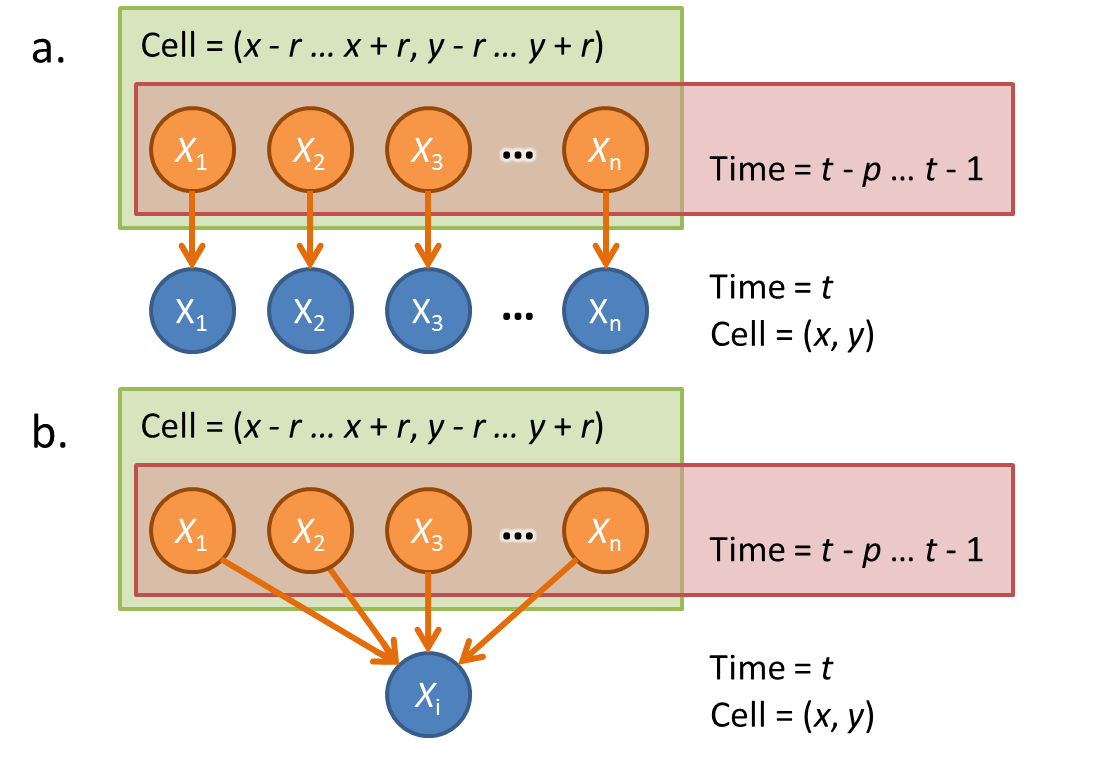


Figure 2. Variables used for estimating values ***X*** at time *t* at cell (*x*, *y*); a. Independent variable model; b. Model with interdependencies between variables.

**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

*λ*(1) is the Lagrange multiplier on the lasso constraints. This is a convex optimization and should be solvable using standard tools.

**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 relative to the actual historical observation ***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**

The region around Pittsburgh was selected as the test case for the proposed methods. 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.

**3.1 Linear Regression**

Four linear regression models were computed for each one of the meteorological values in NLDAS-2. The first two were done 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.

Table 2: Relative testing errors of the linear regression models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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%** |

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 the training and the testing errors, so we can conclude that the models are not over-fitting the data.

**5 Conclusions**

Conclusions

**References**

1. Shuttleworth, J. (2012), *Terrestrial Hydrometeorology*, Wiley-Blackwell.
2. Hong, W. C. (2008), “Rainfall forecasting by technological machine learning methods”. *Applied Mathematics and Computation* **200**(1): 41-57.
3. French, M. N., Krajewsky, W. F. & Cuykendall, R. R. (1992), “Rainfall forecasting in space and time using a neural network”. *Journal of Hydrology* **137**(1): 1-31.
4. Maier, H. R. & Dandy, G. C. (2000), “Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications”. *Environmental Modelling & Software* **15**: 101-124.
5. Zivot, E. and Wang, J. (2006), Vector autoregressive models for multivariate time series”, in *Modeling Financial Time Series with S-PLUS*, pages 385–429. Springer New York.
6. Harris, D., Foufoula-Georgiou, E., Droegemeier, K.K. & Levit, J. J. (2001), “Multiscale Statistical Properties of a High-Resolution Precipitation Forecast”. Journal of Hydrometeorology 2(4):406-418.
7. Hong, W.C. (2008), “Rainfall forecasting by technological machine learning methods”. Applied Mathematics and Computation 200(1): 41-57.
8. Paras & Sanjay, M. (2012), “A Simple Weather Forecasting Model Using Mathematical Regression”. Indian Research Journal of Extension Education 1:161-168.
9. Ferraro, R. R. & Marks, G. F. (1995), “The Development of SSM/I Rain-Rate Retrieval Algorithms Using Ground-Based Radar Measurements”. Journal of Atmospheric and Oceanic Technology 12(4):755-770.
10. Yang, Y., Lin, H., Guo, Z. & Jiang, J. (2007), “A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis”. Computers and Geosciences 33(1): 20-30.
11. National Aeronautics and Space Administration, “Land Data Assimilation Systems (LDAS)”, available online at <http://ldas.gsfc.nasa.gov/index.php>.