CS 229: Final Paper

Wind Prediction: Physical model improvement through support vector regression

Daniel Bejarano (dbejarano@stanford.edu), Adriano Quiroga (aquiroga@stanford.edu)
December 2013, Stanford University

Introduction

The accuracy improvement of weather pattern prediction is in the best interest of humanity. By doing so, we can potentially reduce the impact caused by natural disasters, assess the effects of environmental air pollution in a more precise manner and improve energy dispatch of renewable technologies, particularly wind. The latter goal is the focus of this report, that is, to reduce day ahead wind forecasting error. The use of machine learning (ML) techniques to improve the performance of the weather research and forecasting (WRF) model is explored, with an emphasis on the support vector regression (SVR) method and its nu-SVR variant. The literature pertaining to the use of SVMs for wind prediction is extensive [1] [2] [3], however, as of today there seems to be no incorporation of the WRF model.

The training and testing of the algorithm is done using wind data from an anemometer placed at the Berkeley Yacht club in Oakland CA, spanning the months of October and November 2013, as well as, WRF model outputs for the same time periods provided by Mike Dvorak of Sailor's Energy. The analysis is performed between 18:00PM and 00:00AM and uses the meteorological u-v convention to characterize wind, which allows the model to incorporate speed and direction. All the computer code is written in Python and the ML routines are implementing using the Machine Learning Python (mlpy) module available open source [4].

When dealing with electricity generation and distribution of wind turbines, the National Renewable Energy Lab estimates that a 10% reduction in forecast error translates to about a $140M reduction in grid operating costs [5]. Given the high capital costs associated with current storage technologies, such as pump hydro or batteries, forecast improvement provides a more cost effective way of fully realizing the potential of wind energy. Moreover, by reducing the uncertainty of wind generation, several other issues related to grid operation would be simultaneously addressed, for example, grid instability due to random voltage swings, economic dispatch of generated electricity and unit commitment of generators.

This report is structured into 4 sections, the first provides an overview of the Advance Research WRF model, the second serves provides the reasoning behind choosing SVMs for this analysis, the third presents the results obtained and shows the comparison between the different simulation scenarios, and the last one contains the conclusion and future work.

I. The Weather and Research Forecast Model

It is well recognized and demonstrated that due to the complexity of the system, weather prediction at all time scales is only feasible with numerical models. These models are formulated as initial value problems, that is, they are built using partial differential equations coupled with initial conditions and then solved using numerical methods. The need for starting and boundary conditions turns observational data into one of the pillars of

forecasting. These weather variables are selected from the appropriate sources and desired time scales according to the relevant application. After sorting the relevant data it is further necessary to assimilate it, which implies comparing the observations with model outputs to ensure consistency with expected results from physical laws and adjacent points therefore reducing susceptibility to corrupted data. It is then possible to compute an initial state of the atmosphere, which is then fed as an input to the model.

One such model employed for numerical weather prediction (NWP) is the aforementioned Weather Research and Forecasting (WRF) model. Developed in the late 1990's by collaboration of several government agencies, it is a mesoscale forecasting model designed to cater both to atmospheric research and operational forecasting applications. Furthermore, it features a data assimilation system and allows the user to integrate different dynamic cores and physics packages, making it a versatile forecasting tool [6]. For the purposes of this report, the focus is on wind prediction. The available outputs of the WRF model come from two runs, one using initial conditions from the night before, and the other from the morning of the prediction days. Time resolution is 30 minutes.

The first step of a model run is the WPS, which defines the domains of the simulation, interpolates terrestrial data to fit the domains, and processes meteorological data to fit the grid. In essence, the WPS creates the grid with a given resolution, 200m in this particular case, and discretizes all the appropriate variables to their respective points. The data used for wind prediction is obtained from the National Center for Environmental Prediction and the National Oceanic and Atmospheric administration, and includes, among others, humidity, temperature, pressure and wind values. The second step is data assimilation, although in this particular instance this phase is skipped. The last step is the actual numerical partial differential equation solver. The governing equations of the model include terrain-following hydrostatic-pressure vertical coordinates, moist flux-form Euler equations, map projection, coriolis and curvature terms, all formulated with added perturbation variables to reduce truncation and machine rounding errors during the numerical calculations. The ARW solver uses a discrete time integration scheme, incorporating several schemes, third-order Runge-Kutta, forward-backward and vertically implicit [7].

II. SVR method

Wind speeds are typically modeled using the Rayleigh distribution [8]. From figure 1 it is possible to distinguish this distribution for the October/November 2013 data period both throughout the hours of a day, and for each individual hour. This tendency of wind to vary between certain values with a particular probability is what motivated us to use supervised learning methods with forecasts made each half hour. Several case studies analyzed concluded that SVMs provide lower error compared to other approaches, especially neural-networks [1] [9]. As progress was made it became obvious that the results to our analysis would be very similar to those obtained from previous studies of the same nature. Therefore, to increase the robustness of the analysis and with the intent of contributing a new approach to the literature, we decided to use Support Vector Regression and to incorporate the WRF predictions into it.

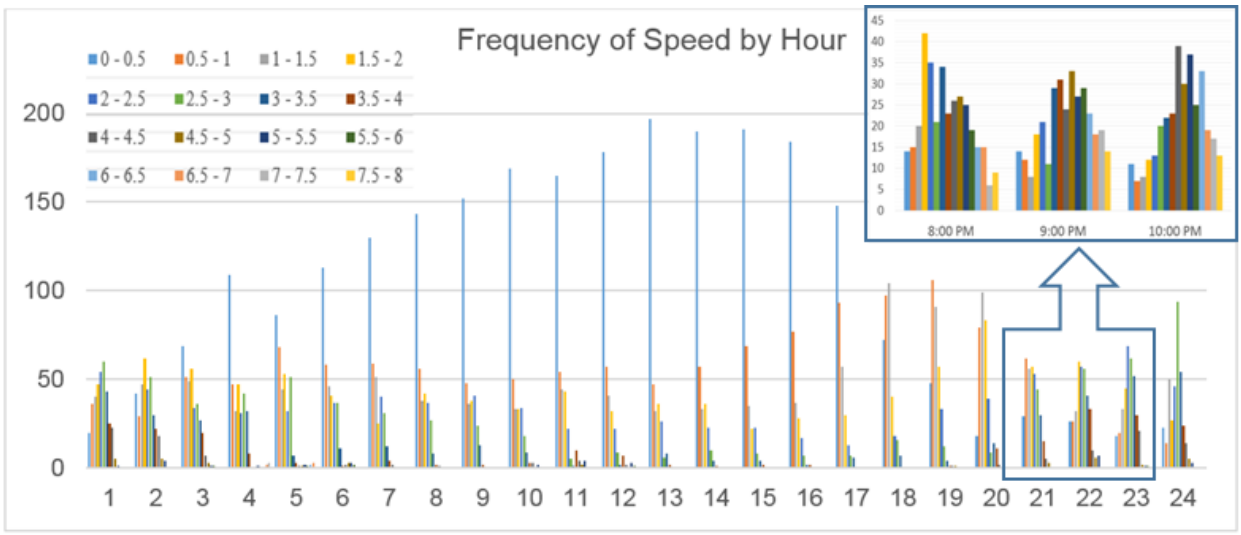

Fig. 1: Wind speed frequency distribution by hour. Resembles the Rayleigh distribution both on a daily and hourly basis.

III. Simulation Setup and Results

The simulations are structured as follows, wind speed data from a fixed number of days before the prediction day is used as the training set, with the day immediately after set as the training prediction. For example, the algorithm takes 14 days to recognize a pattern and the 15th as the desired output of the prediction, for which data has already been gathered. With the regression already performed and the SVs extracted, the code then predicts wind outputs for the training prediction day and one day after, in the previous scenario the 15th and 16th days. In a rolling window fashion, the algorithm then performs several train-predict iterations to obtain wind speed estimates for the entire month. The forecasts are made each half hour interval, corresponding to the available resolution of the WRF outputs. The model performance is characterized using the root-squared-error (RMSE) summing both u and v components.

The performance of the SVR method was studied across different scenarios varying the number of days used for training, the cost of constraint violation parameter (C'), the nu-parameter (v) and most notably, the types of data used for the regression, three in the case of this analysis. They are, real measurements from the Berkeley Yacht Club anemometer, WRF model outputs at the location closest to the anemometer and a mixture of both. The idea behind this type of implementation is to explore WRF error prone time instances and try to improve accuracy. The results of the forecast for the entire month of November 2013 at 6PM are shown in figure 2. It is clear from the graphs that the WRF model predictions exhibit the greatest variance, while the ML ones tend to stay close to the mean. However, since wind speeds are distributed according to Rayleigh distributions [8], the latter behavior is generally more desirable. In the context of the ML algorithms, this implies the estimates will be more accurate if outliers are ignored during training, which is

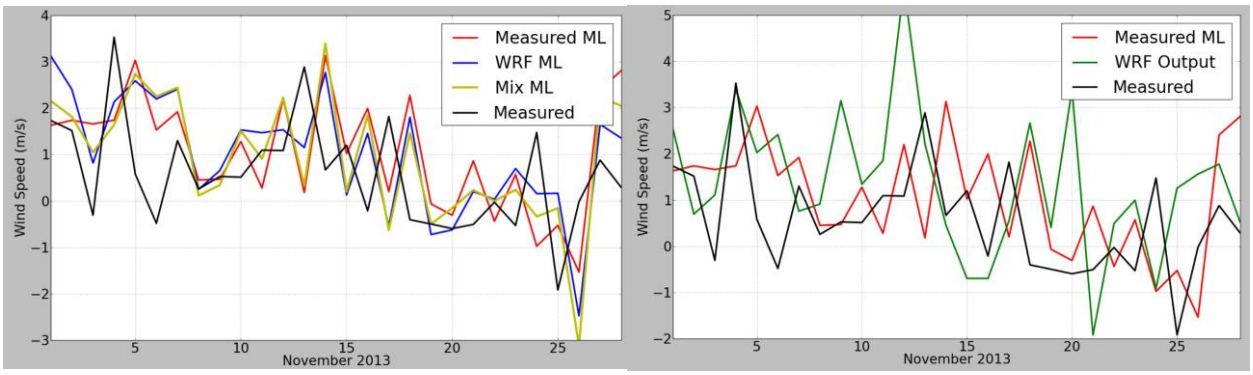

Fig. 2: Measured and forecasted wind speed values for the entire month of November 2013, compared across different training data (left) and WRF model outputs (right).

confirmed by figure 3, which shows an RMSE increase as the cost of constraint violation parameter is increased, i.e., as the regression gets tighter [10].

Also in figure 3, it is possible to see the effect of changing the number of training days on the RMSE. Taking a larger number of them causes the same error as taking only few of them, suggesting there is an optimal number of training days. However, this is dependent on the type of data used for the predictions, around 10 days in the case of measured speeds and 14 days in the case of WRF outputs. This optimal point is most likely a function of the forecast period as well. Moving forward, it will be important to determine whether the RMSE improvements are worth optimizing these values.

Figure 4 shows the forecast for November 20th 2013. The results show the potential of incorporating other meteorological variables to the predictions. At 20:30 we see that the mixed data is closer to the actual speeds than the two others. By incorporating physical models of the environment through the WRF, the training takes into account more than just previous wind patterns, correlating them with temperature, humidity and pressure values as well. However, looking at 22:00, error reduction is not always guaranteed. Hence, further study is needed to determine whether it is more beneficial to incorporate

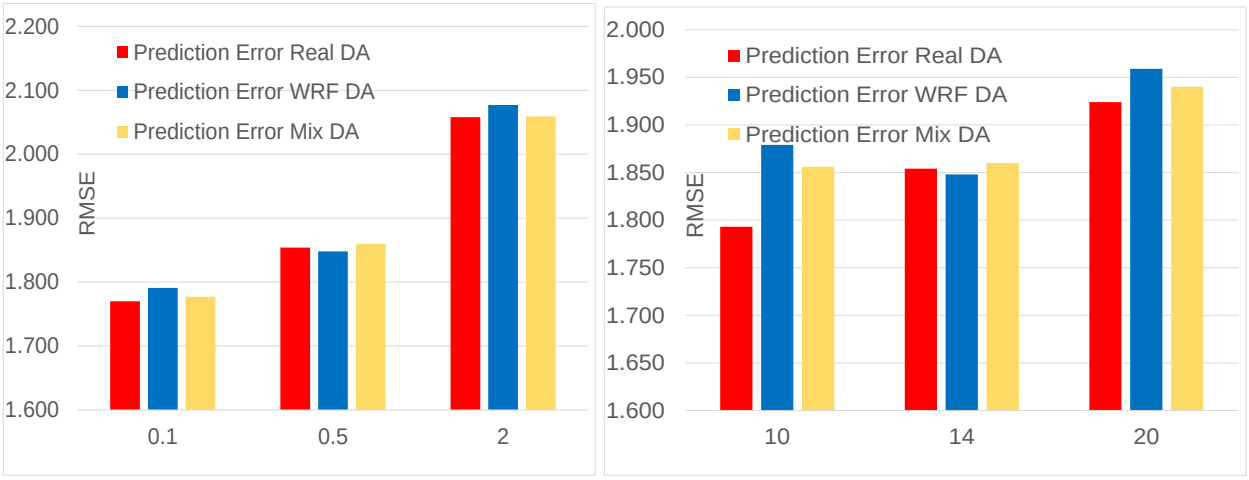

Fig. 3: RMSE values for the three different types of training data, varying cost of constraint violation (left) and number of training days (right).

other types of data directly, that is, feed them straight to the SVR algorithm instead of using the WRF.

Finally, the RMSE for the WRF is **2.417**, which means the SVR method can outperform the WRF model by as much as **20%** for the Berkeley Yacht club case.

IV. Conclusions and future work

Wind speed prediction through the combination of support vector regression and the weather research and forecast model was explored. This ML rolling window method can outperform the WRF in the case of the Berkeley Yacht Club data. Interestingly, the amount of previous day variables needed for this degree of accuracy do not exceed 15 days.

As this approach is developed, there are some issues that require additional exploration. First, validate the results with datasets of greater resolution, not only in time, but also with a larger amount of points in space. Second, incorporate other weather variables, such as temperature humidity and pressure directly and analyze up to which point the inclusions are beneficial. Third, determine what causes the unusual peaks on the WRF model, figure 2 around day 13 of the green line. Fourth, perform time series analysis to better understand time correlations and potentially be able to predict wind ramps. Finally, explore the possibility of incorporating a probabilistic model directly into the forecast. There is still room for improvement.

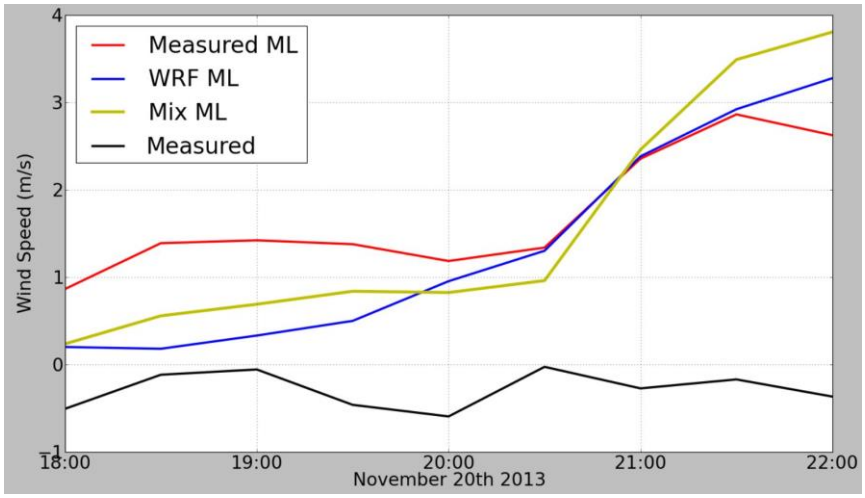

Fig. 4: Measured and forecasted u component for November 20 compared across the three different types of training data. At 20:30 the mix data is closer to the actual value, the opposite is the case at 22:00.

Acknowledgements

The authors would like to thank Mike Dvorak of Sailor's Energy for providing the necessary WRF output values, the computing resources to run the simulations and for his support in the development of this project.

Bibliography

[1] M.A. Mohandes, T.O. Halawani, S. Rehman and A.A. Hussain, "Support vector machines for wind speed prediction," J. Renewable Energy, **29** [2004] 939-947

[2] P.J. Sallis, W. Claster and S. Hernandez, “A machine-learning algorithm for wind gust prediction,” J. Computers Geosciences, **37** [2011] 1337-1344

[3] M. Bhaskar, A. Jain and N.V. Srinath, “Wind Speed Forecasting: Present Status,” Int. Conf. Power System Tech. [2010] 1-6

[4] D. Albanese, “mlpy Documentation, Release 3.5,” March 2012

[5] R. Piwko and G. Jordan, “Impacts of Improved Day-Ahead Wind Forecasts on Power Grid Operations,” NREL Report [2010]

[6] Advanced Research WRF user’s guide, http://www.mmm.ucar.edu/wrf/users/docs/user_guide_V3/contents.html

[7] Technical description of the Advanced Research WRF, http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf

[8] F.M. Vanek, L.D. Albright and L.T. Angenent, “Energy Systems Engineering: Evaluation and Implementation,” McGraw Hill, New York USA, 2012

[9] P. SangitaB and S.R. Deshmukh, “Use of Support Vector Machine for Wind Speed Prediction,” Int. Conf. Power Energy Systems, [2011] 1-8

[10] C.C. Chang and C.J. Lin, “LIBSVM: A Library for Support Vector Machines” http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf