Energy Research and Development Division

FINAL PROJECT REPORT

Improving Short-Term Wind Power Forecasting Through Measurements and Modeling of the Tehachapi Wind Resource Area

California Energy Commission

Edmund G. Brown Jr., Governor

February 2018 | CEC-500-2018-002

PREPARED BY:

Primary Authors:

Aubryn Cooperman
C.P. van Dam
John Zack
Shu-Hua Chen
Clinton MacDonald

California Wind Energy Collaborative
Mechanical and Aerospace Engineering Department
University of California, Davis
One Shields Avenue
Davis, CA 95616
cwec.ucdavis.edu

Contract Number: EPC-14-007

PREPARED FOR:
California Energy Commission

Mike Kane
Project Manager

Aleecia Gutierrez
Office Manager
ENERGY GENERATION RESEARCH OFFICE

Laurie ten Hope
Deputy Director
ENERGY RESEARCH AND DEVELOPMENT DIVISION

Drew Bohan
Executive Director

ACKNOWLEDGEMENTS

The authors acknowledge the helpful comments and insight of Amber Motley and Jim Blatchford of the California Independent System Operator. The authors also thank the Kern County Public Works Department, the National Chavez Center, Computronics, and EDF Renewable Energy for permitting the installation of measurement equipment on their properties for the duration of the research, as well as Earth Networks for providing data.

PREFACE

The California Energy Commission's Energy Research and Development Division supports energy research and development programs to spur innovation in energy efficiency, renewable energy and advanced clean generation, energy-related environmental protection, energy transmission and distribution and transportation.

In 2012, the Electric Program Investment Charge (EPIC) was established by the California Public Utilities Commission to fund public investments in research to create and advance new energy solution, foster regional innovation and bring ideas from the lab to the marketplace. The California Energy Commission and the state's three largest investor-owned utilities – Pacific Gas and Electric Company, San Diego Gas & Electric Company, and Southern California Edison Company – were selected to administer the EPIC funds and advance novel technologies, tools, and strategies that provide benefits to their electric ratepayers.

The Energy Commission is committed to ensuring public participation in its research and development programs that promote greater reliability, lower costs, and increase safety for the California electric ratepayer and include:

- Providing societal benefits.
- Reducing greenhouse gas emission in the electricity sector at the lowest possible cost.
- Supporting California's loading order to meet energy needs first with energy efficiency and demand response, next with renewable energy (distributed generation and utility scale), and finally with clean, conventional electricity supply.
- Supporting low-emission vehicles and transportation.
- Providing economic development.
- Using ratepayer funds efficiently.

Improving Short-Term Wind Power Forecasting Through Measurements and Modeling of the Tehachapi Wind Resource Area is the final report for the Advanced Modeling and Measurements in Tehachapi project (Contract Number EPC-14-007) conducted by the California Wind Energy Collaborative. The information from this project contributes to Energy Research and Development Division's EPIC Program.

All figures and tables are the work of the author(s) for this project unless otherwise cited or credited.

For more information about the Energy Research and Development Division, please visit the Energy Commission's website at www.energy.ca.gov/research/ or contact the Energy Commission at 916-327-1551.

ABSTRACT

This report describes atmospheric measurements and modeling in and around the Tehachapi Wind Resource Area to improve wind power forecasting in the short term (0-15 hour ahead) and very short term (0-3 hours ahead). The measurement component of the project involved maintaining and expanding a network of sensors targeted at improving wind speed forecasts in the Tehachapi Wind Resource Area. When combined with earlier measurements, this network produced a 25-month dataset of wind speeds, temperature, humidity, and other relevant parameters in the Tehachapi Pass. The modeling component used mechanical learning techniques to develop predictors for very short-term forecasts. An initial study of model configurations led to the selection of an appropriate set of submodels for the Tehachapi Wind Resource Area. These submodels, which included a wind turbine drag submodel, had a significant effect on the mean absolute error and bias of the wind power forecasts. A machine learning method was used to identify predictors for the 15-minute average power production, 60-minute ramp rates, and occurrence of large ramp events in the 0-3 hour look-ahead period. The forecasting improvements developed as part of this project, including data from project sensors, improved data assimilation methods, and machine-learning-based predictors were combined in the improved operational forecast system. In a six-month forecast evaluation, the improved system achieved a 13.5 percent reduction in the mean absolute error of the 15-minute average power production in the Tehachapi Wind Resource Area against an optimized ensemble of National Weather Service forecast models.

Keywords: *wind power forecasting; wind ramp events; atmospheric sensors; numerical weather prediction; data assimilation; machine learning;*

Please use the following citation for this report:

Cooperman, Aubryn, C.P. van Dam, John Zack, Shu-Hua Chen and Clinton MacDonald. 2018. *Improving Short-Term Wind Power Forecasting Through Measurements and Modeling of the Tehachapi Wind Resource Area*. California Energy Commission. Publication Number: CEC-500-2018-002.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Introduction

As installed wind power capacity in California grows, reliable forecasting of wind power production becomes essential. A significant challenge for forecasters is predicting wind ramps - large increases or decreases in wind speed over a short period. When a wind ramp occurs in an area with significant wind energy generation capacity, it can cause a disruptive change in the amount of power delivered to the grid. Uncertainty in the short-term prediction of wind power production leads grid operators to require more reserve power generation, which may be more expensive or emit more pollutants than wind turbines. Improving wind power forecasting reduces the uncertainty associated with wind power production and contributes to improved grid operations.

Project Purpose

This project explored improving wind power forecasting in the Tehachapi Wind Resource Area in Kern County; a source of more than half of California's annual wind energy production. A large wind ramp event in the Tehachapi Wind Resource Area can reduce power output by 1,000 megawatts (MW) or more within an hour. The Tehachapi Wind Resource Area is a major contributor to California's energy supply and has been the focus of previous studies, which allowed this project to take advantage of an existing set of meteorological instruments.

Project Process

This project used meteorological sensors to detect and measure precursors to wind ramps and developed more advanced modeling tools to forecast wind power production. The atmospheric sensors previously installed, as part of an earlier study, were operated over a longer period, while additional sensors expanded the range of instruments providing data. These sensors measured wind speed and direction, temperature, humidity, and cloud height at four sites in and upwind of the Tehachapi Wind Resource Area.

The modeling component of this project took two approaches to improving wind ramp forecasts: numerical weather prediction and empirical modeling. Numerical weather prediction models attempted to simulate the physical processes occurring in the atmosphere. The National Oceanic and Atmospheric Administration develops and runs several of these models, which are used by forecasters such as the National Weather Service and for industry-specific forecasts including aviation, wind and solar energy, and others. The models use data from weather sensors throughout the United States, and the model outputs, code, and data are publicly available.

This project focused on improving the ability of the Weather Research and Forecasting model to predict wind ramp events in the Tehachapi Wind Resource Area. The team identified the best Weather Research and Forecasting model configuration for a sample of 30 ramp events, and then used that configuration as a baseline to evaluate the impact of project sensor data and model improvements. These changes to the Weather Research and Forecasting model configuration aimed to improve turbine-height wind forecasts and ramp event forecasts in the Tehachapi Wind Resource Area by assimilating additional observations from project sensors and using high-spatial-resolution topography and land cover datasets.

Empirical models for predicting wind ramps do not attempt to reproduce the physical mechanisms that occur in the atmosphere to produce changes in wind speed; instead, they identify statistical correlations between weather conditions at different locations and times. For example, a weather station upwind of a wind power plant might consistently observe an increase in wind speed before wind plant operators see an increase in power production. The challenge for modelers is to identify predictors that will reliably indicate upcoming ramps without excessive false positives or missed events.

The very short-term (0-3 hours) forecast component of this project compared several machine learning approaches to identify wind ramp predictors using a year of data from the project sensors and other publicly available weather measurements in the Tehachapi region. Based on these experiments, the Extreme Gradient Boosting method was selected for the very short-term forecast system designed to produce updated power production and wind ramp forecasts every 15 minutes.

The final product of the modeling effort was an Improved Operational Forecast System that incorporated the Weather Research and Forecasting improvements and the statistical correlations identified during this research. The Improved Operational Forecast System was assessed as part of a multimethod forecast system that represents current state-of-the art methods used for wind power forecasting.

Project Results

Model Sensitivity Experiments

The model sensitivity experiments compared 11 configurations of the Weather Research and Forecasting model for 30 Tehachapi Wind Resource Area wind ramp events in late 2014 and 2015. Results from the selected model configurations matched the observed wind speeds with varying degrees of accuracy. An important factor affecting the results was whether the model accounted for the presence of wind turbines. When turbines extract energy from the wind, they slow wind speeds downstream. The slowdown is most pronounced at the turbine hub height, which is also the most important height for wind power forecasting. At the scale of a regional weather forecast, the impact of a single turbine is minimal and would not typically be modeled, but the hundreds of turbines in the Tehachapi area have a cumulative effect. All the model configurations that did not include the turbines predicted much higher wind speeds than actually occurred, while the models that included the impacts of turbines on wind speed matched observations more closely. Counterintuitively, including turbines in the model led to worse prediction of large ramp events, most likely because the slower wind speeds meant that some changes in wind speed did not meet the threshold to be counted as ramps. A configuration including wind turbines in the model was selected as the baseline for the remaining forecast improvement experiments.

Short-Term Wind Ramp Forecasting Improvements

Two studies were conducted to examine the effect of Weather Research and Forecasting model improvements on forecasts of wind speed at turbine hub height (80 meters) and wind ramp forecasts. The first compared forecasts over one month, while the second focused on 10 selected ramps. Both studies compared different strategies for incorporating sensor data into the numerical weather model.

Using project sensor data improved wind ramp forecasts for the Tehachapi Wind Resource Area. The effect of sensor data on wind speed forecasts depended on the location where forecasts were evaluated and the frequency of data updates. Forecasts that reduced the wind speed error at one measurement site increased the error at another site. The effects of assimilating sensor data were seen mainly in the first six hours, with the largest effects in the first hour.

Very Short-Term Wind Ramp Forecasting Improvements

The very short-term forecasts attempted to answer three questions about the upcoming three-hour period. What will the average wind speed be (in blocks of 15 minutes)? What will be the steepest increase or decrease in wind speed? Will there be a large wind ramp?

The forecasts were evaluated by comparison with a persistence forecast (in other words, the wind speed does not change from the time the forecast is issued). The error in predicting average wind speed was reduced by more than 25 percent, while errors in the wind speed rate of change were reduced by more than 30 percent. Because a persistence forecast will never predict a ramp event, predicting large wind ramps was evaluated using the Critical Success Index, which is the ratio of correctly forecasted events, or "hits", to the total number of hits, misses, and false alarms. The forecasts scored around 30 percent on this index for ramps of more than 300 megawatts (MW), with decreasing accuracy at predicting the less-frequent, larger ramps. Comparison of forecasts with and without project sensor data indicated that the data improved forecast performance, in particular for the prediction of the largest ramps.

Overall, the machine-learning-based very short-term prediction model exhibited considerable skill in the 0-3 hour prediction of the time series of the 15-minute average power production, the maximum and minimum ramp rates and the occurrence/nonoccurrence of large ramps. The project sensor data contributed to substantial improvements in the performance of all three forecast modes.

Improved Operational Forecast System

The research team evaluated the improved operational forecast system on the basis of forecasts for a period of six months in 2015. Output from the improved system was combined with output from three National Weather Service models to produce an ensemble forecast. A second ensemble forecast using only the National Weather Service models was used as the baseline for comparison. The Improved Operational Forecast System reduced the error in the forecast of power production in the Tehachapi Wind Resource Area by 13.5 percent. Of that reduction, 6.7 percent was a result of the improvements in the numerical weather prediction model, while the other 6.8 percent was from using very short-term statistical forecasting methods. Most of the improvement was in the first three hours of the 15-hour forecast window, which suggests that the reduction in forecast error was associated with the effective use of local data from the project sensor network.

Benefits to California

This project produced high-quality data from models to help forecasters better predict future wind ramps. The project also highlights the value of a long-term, stable network of meteorological instruments to provide data for improved forecasts. Data from project sensors had a significant impact on forecast skill in the very short term (0-3 hours ahead):

- There was a 7.2 percent reduction in error margins of Tehachapi Wind Resource Area aggregate power production forecast
- There was a 6.9 percent reduction in error margins of ramp rate forecast
- The project improved improved prediction of large ramp events (>750 MW)
- The sensor with the largest impact on very short-term forecasts was a radar wind profiler that measured wind speeds up to 4000 meters above ground level, located upstream of Tehachapi Pass.

This project produced several quantifiable improvements in wind speed forecasting that can be immediately implemented in forecasts provided to the California Independent System Operator, utilities and wind plant operators. Compared to current forecasts, the improved forecasting system reduced the error in the power production forecast for the Tehachapi Wind Resource Area by 13.5 percent. Half of the improvement is attributable to better use of observational data in the forecast model, while the other half is due to the use of machine learning to identify statistical correlations within a three-hour forecast window.

CHAPTER 1: Introduction

Wind plants in California produced 13.5 trillion watt-hours (TWh) of electricity in 2016, representing 6.8% of the total in-state power generation.[1] Because wind is a variable resource, the amount of power generated from wind changes frequently. Often, changes are small and a decrease in power output by one turbine may be offset by an increase in power output by another turbine. At other times, the wind speed increases or decreases sharply over a large area in a short period, called a *wind ramp*. Predicting ramps is of significant interest for wind power generators and power grid operators because these events can change power output 1,000 megawatts (MW) or more. Improving the ability to predict significant wind ramps hours in advance allows grid operators to better maintain the balance of generation and load.

This project addresses two elements of improving wind power forecasting: using meteorological sensors to detect precursors to wind ramps, and developing more advanced modeling tools to forecast wind power production. The project focuses on the Tehachapi Wind Resource Area (TWRA) in Kern County, which contains more than half of California's wind generation capacity.

1.1 Overview of Previous Work

This project builds on two previous studies that investigated the selection and siting of sensors for short-term and extreme-event forecasts and installed sensors in Tehachapi.

1.1.1 WindSENSE

The WindSENSE project (Manobianco et al., 2011) was a collaboration between Lawrence Livermore National Laboratory and AWS Truepower, LLC (AWST), funded by the Department of Energy. The project aimed to improve wind power generation forecasting by better predicting large ramps. One outcome of the WindSENSE project was a list identifying locations in the TWRA (among other regions) where installing meteorological sensors would produce the most benefit. These locations were the basis for the instrument installations in the WindSense project.

1.1.2 WindSense

The WindSense project (Kamisky et al., 2016), funded by the California Energy Commission, was carried out by partners from the University of California, Davis (UC Davis), Sonoma Technologies, Inc. (STI), and DNV GL. This project installed measurement instruments at several locations in the TWRA that had been identified as potentially beneficial for wind ramp forecasting and to assess the impact of those sensors on forecasts. The instruments used and

[1]California Energy Commission, 2016. *Total System Electric Generation in Gigawatt Hours*, Nyberg, http://www.energy.ca.gov/almanac/electricity_data/total_system_power.html, June 23, 2017

data collected as part of the WindSense project were also anticipated to contribute to a future, more in-depth forecasting effort such as the project described in this report.

Measurement instruments were installed at five locations in the TWRA:

- At Bena Landfill in the southern San Joaquin Valley, a radar wind profiler and radio acoustic sounding system provided wind and temperature profiles up to 3000 m above ground level (AGL).
- At the National Chavez Center near Keene, a sodar (Sonic Dection And Ranging – a meteorological instrument used as a wind profiler) was used for continuous measurements of winds up to 600 m AGL.
- A second sodar was installed at a site called EDF Avalon in the Mojave Desert. This site is within a wind plant.
- At the Windmatic site near Tehachapi Airport, two instruments were installed: a mini-sodar capable of measuring winds up to 200 m AGL and a microwave radiometer for measuring temperature, humidity, and liquid water profiles.
- A second microwave radiometer was located at the Bakersfield Airport.

These instruments collected data from March 1, 2015 to August 31, 2015. The data were displayed on a project website, distributed to modelers, and archived for future use. Eleven significant ramps were identified during the seven-month period. A numerical weather prediction (NWP) model (WRF) and artificial neural network were used to generate forecasts with data from various combinations of the project sensors, as well as without any project data, to assess the impact of the remote sensing technology. No single set of measurements was found to consistently benefit the wind power forecast in this study. Recommendations from the WindSense project included lengthening the study period to include the windy season and investigating alternative configurations of the NWP model.

1.2 Project Outline

1.2.1 Project Goals and Objectives

This project aimed to leverage the instrumentation, recorded data, and experience gained from the previous projects to improve wind ramp forecasting in the TWRA, focusing on short-term (0-15 hour) and very-short-term (0-3 hour) forecasts. The objectives were to:

- Complete a forecast sensitivity error analysis to identify and quantify the parameters that most significantly affect wind ramp forecast errors.
- Conduct a one-year measurement campaign in the TWRA, focused on the phenomena that drive wind ramps.
- Implement improvements to computational modeling of flow physics at low levels in complex terrain.
- Implement statistical and empirical methods to make very short-term correlations between meteorological measurements and wind turbine and wind plant production.

- Incorporate the improvements to computational modeling and the statistical and empirical correlations described above into a state-of-the-art wind power forecast system.
- Validate the modeling improvements for low levels in complex terrain and immediately incorporate them into forecasts of wind power and wind power ramps in the TWRA provided to the California Independent System Operator (ISO).

1.2.2 Project Tasks

The project was divided into five technical tasks, each discussed in the following chapters.

- Task 2: Model Sensitivity Experiments (Chapter 2): The goal of this task was to determine the sensitivity of wind ramp forecast errors to key components of WRF model physics schemes.
- Task 3: Field Measurement (Chapter 3): This task entailed operating remote sensing instruments—those installed during the WindSense project as well as new sensors—for one year.
- Task 4: Short-Term Wind Ramp Forecasting Improvements (Chapter 4): This task aimed to improve critical components in the short-term forecasting of wind ramps, including data assimilation, planetary boundary layer parameterization, and spatial resolution.
- Task 5: Very-Short-Term Wind Ramp Forecasting Improvements (Chapter 5): This task focused on the use of empirical and statistical methods to identify precursors in data collected by project sensors and other sources to improve the prediction of wind ramps in the 0-3 hour-ahead time frame.
- Task 6: Wind Ramp Forecast System Evaluation (Chapter 6): The goal of this task was to configure, operate, and evaluate the real-time Baseline Operational Forecast System (BOFS), real-time Enhanced Baseline Operational Forecast System (EBOFS), and retrospective Improved Operational Forecast System (IOFS).

CHAPTER 2: Model Sensitivity Experiments

2.1 Introduction

The experiments in this component of the project were designed to assess the sensitivity of forecasts for significant wind ramp events in the TWRA to the WRF model configuration. The total wind power generation capacity is more than 3,000 megawatts (MW). However, only 17 facilities with an aggregate, or combined, capacity of 2,319 MW were considered in the experiments in this project. These were the facilities that had provided a substantial period of high-quality data to the California Independent System Operator (California ISO) for wind forecasting applications.

2.2 Design of Sensitivity Experiments

2.2.1 Baseline Configuration

The baseline configuration of the WRF modeling system used in these experiments was a configuration that has been widely used for short-term forecasting in California and elsewhere by several private and public entities. It has been used for wind prediction (Deppe et al. 2013), real-time fire weather (CANSAC 2015) and air quality applications (Rogers et al. 2013) (Figure 1). The specifications of the key submodels for this configuration are provided in Appendix A.

Figure 1: Geographical Domains of the WRF Grids Used in the Sensitivity Experiments

Nested WRF model grid configuration (left panel) used for the sensitivity case studies. Inner (outer) domain has grid spacing of 1 km (3 km). Right panel shows expanded view of 1-km domain with field instrument locations (blue boxes), wind plant sub-aggregate sites (red circles), and Automated Surface Observing System weather stations (yellow stars).

The forecasts from the baseline ran as well as those from all of the alternate configurations were initialized with data from the US National Weather Service's Rapid Refresh (RAP) model

(Benjamin et al, 2016). The experimental WRF forecasts were started roughly six hours before the observed ramp event in each case and had a forecast length of 15 hours.

2.2.2 Alternate Configurations

Once the baseline simulation (P0) was completed, a set of experimental runs was executed to investigate the sensitivity of the WRF forecasts to the submodels that are used. The configuration of the model for each experiment is listed in Appendix A.

Experiment P1 was designed to test an alternate surface boundary layer submodel set. The WRF configurations in experiments P2 through P5 have been used in previous studies for air quality (Hu et al. 2010) and other applications. Each of these experiments changes only one submodel in the P0 configuration. Experiment P6 tested the use of higher resolution data sets to initialize terrain and surface properties.

Experiments P7, P8 and P9 were designed to test the effect of using a wind turbine parameterization scheme that accounts for the effects of wind turbines as momentum sinks on the mean flow while increasing turbulent kinetic energy in the model in the lowest model layers containing rotors (Fitch et al. 2012). These experiments were motivated by the fact that metrics presented in later sections of this report show a high 80-meter wind speed bias for the P0 to P6 experiments over the entire TWRA where the model has no indication of existing turbines. This result suggested that such a parameterization might improve the forecasts especially in regions with a high turbine density.

In addition to the experiments (P0-P9) that tested using different physics submodels, two additional experiments (R1 and R2) were conducted to test the effect of changing the horizontal grid resolution. Unfortunately, using the R2 configuration at a resolution of 1 km caused instability issues in the WRF model resulting in aborted runs for several cases in the experimental sample. As a result, the statistical results for the R2 configuration were based on only 16 of the 30 cases in the experimental sample.

2.2.3 Case Sample Composition

The experimental sample consisted of 30 cases with the largest 60-minute ramp rates (change in power generation per minute) during a recent 1.5-year period, a coherent ramp feature in the generation time series and a good availability of observational data. Cases with smaller ramp rates, a noisy temporal evolution, or lack of observational data were generally excluded.

There were five up and five down ramps (i.e., 10 cases) selected from each of the three major TWRA weather regimes: (1) diurnal cycles (May – July), (2) monsoonal flows (August – September) and (3) midlatitude events (December – February) from August 2014 –June 2015.

2.3 Results of Experiments

The experimental WRF forecasts were evaluated from four perspectives: (1) the time series of the average 15-minute hub-height wind speed and power production for the TWRA aggregate; (2) ramp events for the TWRA aggregate, which were defined as the exceedance of a specified threshold for the 60-minute ramp rate; (3) the vertical profile of wind speed at three of the project sensor sites and (4) the spatial and temporal evolution of selected ramps.

2.3.1 Time Series Forecasts

The time series forecasts consisted of the prediction of the 15-minute average power generation for the TWRA aggregate (2,319 MW) and six subaggregates as well as the 15-minute capacity-weighted hub-height wind speed for the 0-15 hour forecast period. The aggregate power generation was calculated by summing the forecasted power production for each wind generation facility.

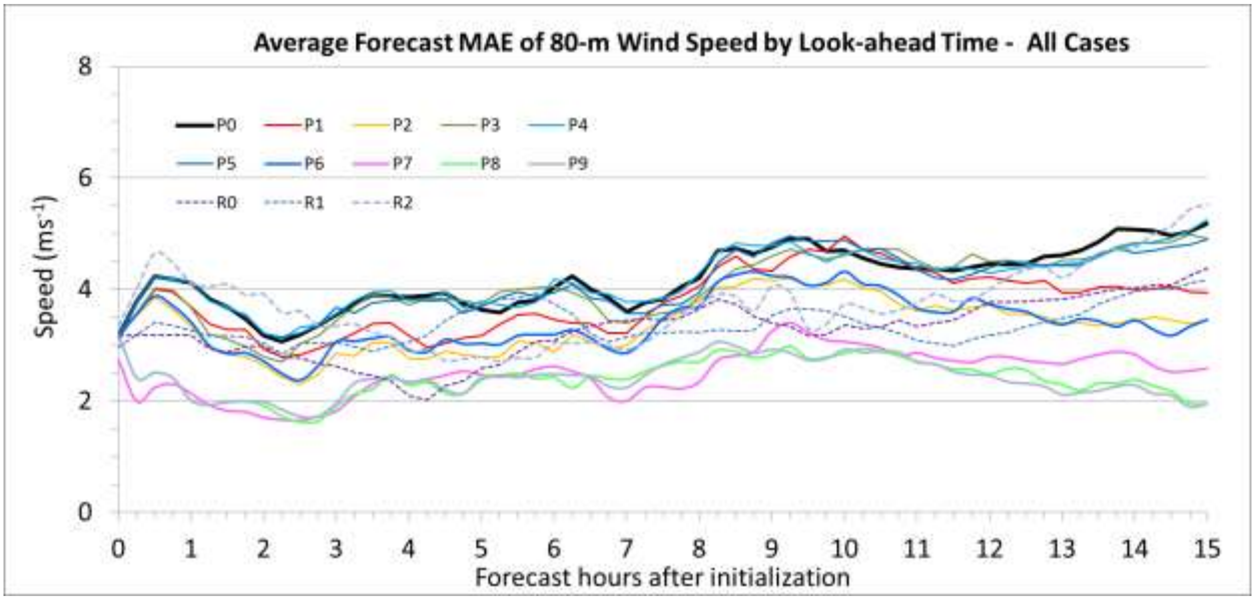
This evaluation of the time series forecasts was based on three standard performance metrics: (1) mean error (bias), (2) mean absolute error (MAE), and (3) root mean square error (RMSE). These metrics were calculated by look-ahead time in 15-minute intervals for the power production and wind speed forecasts and compiled for the entire 15-hour forecast period.

Charts of the bias and MAE of the aggregate capacity-weighted wind speed forecasts by look-ahead time are depicted in Figure 2 and 3. An examination of these charts indicates several key points:

(1) The wind speed and power predictions from all of the experimental configurations had a substantial positive bias except for the configurations that employ the turbine drag submodel (P7, P8 and P9).
(2) The baseline configuration (P0) had the highest bias, MAE and RMSE,
(3) The configurations (P7, P8 and P9) that employ the turbine parameterization had the lowest bias, MAE and RMSE.
(4) Among the parameter configurations that did not employ the turbine drag submodel, the P2 configuration had the lowest bias, MAE and RMSE,
(5) The NWS RAP model (R0) at 13-km grid resolution produced bias, MAE and RMSE values that were similar to the best high resolution (1 km) WRF configuration without the turbine drag submodel (P2).
(6) The high resolution (1 km) version of the RAP configuration (R2) had slightly higher MAE and RMSE values for the wind speed forecasts than the lower resolution (13 km) configuration but the higher resolution version achieved slightly lower MAE and RMSE for the power production forecasts.

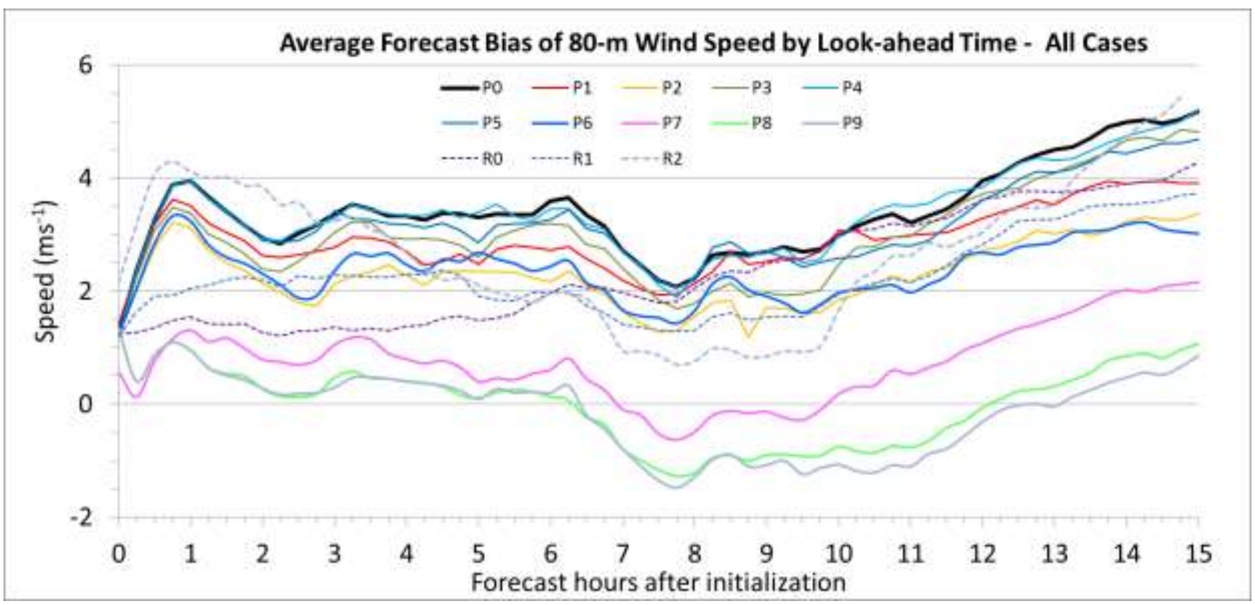
It is clearly evident that there is a general increase in MAE and bias over the 15-hour forecast period with the MAE or bias being higher in the later hours of the forecast period as would be expected (i.e., the error grows over look-ahead time). However, for most of the experiments the increase in MAE is actually concentrated in the six to nine-hour look ahead period (Figure 2). The MAE is fairly flat before this period and is either fairly flat or slightly decreasing after this period. As noted earlier, the initialization times of the simulations were selected such that the ramps occurred about six to nine hours after the initialization time. Thus, the significant increase in MAE occurs during the general time periods of the ramps. Interestingly, the wind speed bias decreases significantly during this period.

Figure 2: MAE of Wind Speed Forecasts by Look-Ahead Time for Each Sensitivity Experiment



Mean absolute error (MAE, m s^{-1}) of 80-m wind speed by look-ahead time for all WRF experiments for the 30 large wind ramps in 2014 and 2015. Note R2 includes only 16 of the 30 cases.

Figure 3: Bias of Wind Speed Forecasts by Look-Ahead Time for Each Sensitivity Experiment



Bias (m s^{-1}) of 80-m wind speed by look-ahead time for all WRF experiments for the 30 large wind ramps in 2014 and 2015. Note R2 includes only 16 of the 30 cases.

Overall, the results indicate that MAE and RMSE scores of the experiments were strongly tied to the magnitude of the bias in each configuration. Thus, bias reduction was a major factor in differentiating the performance among the configurations. While lower bias is a desirable attribute, the reduction in bias by itself does not necessarily provide a more valuable NWP forecast. First, the forecast with the lower bias may not have any additional information about the timing or amplitude of the significant events. Second, the bias, if it is consistently present, can be significantly reduced or entirely eliminated through using statistical post processing (e.g., Model Output Statistics as in Glahn and Lowry, 1972) while retaining the other desirable aspects of the NWP forecasts.

Based on the bias, MAE and RMSE alone, the configurations (P7, P8 and P9) that incorporated the wind turbine drag submodel were clearly the best performers. In fact, P8 and P9 were the best performing simulations based on these metrics with no significant performance difference between them. The forecasts from these experiments had near zero bias for the wind speed and a low amplitude bias for the predictions of the wind generation. Thus, the addition of the turbine drag submodel eliminated much of the high bias that was seen in the other experiments. That raised the question of whether the bias in the other simulations was attributable largely to the omission of the turbine drag physics or if it was actually attributable to other factors and the application of the turbine drag physics provided a better answer for the wrong reasons (i.e., it conveniently compensated for other errors in the model physics). The research team addressed this question by evaluating other aspects of the simulations.

2.3.2 Ramp Event Forecasts

The ability to predict the occurrence or nonoccurrence of significant ramp events was also evaluated. The evaluation was based on a deterministic assessment of a correct or incorrect forecast of a ramp event. A ramp event was defined as a net change of about 20% of capacity over a 60-minute period. For the evaluation, a “time window” was defined to permit modest phase errors in the prediction and still count the prediction as correct. Statistics were computed for three time windows with a half-width of 60, 120 and 180 minutes centered on the forecasted or observed event start time. Only the results for the 120-minute statistics are presented in this report. The forecasted events were based upon the power production changes in the time series of the 15-minute average power production.

If an observed ramp occurred within the time window of a predicted event, the prediction was classified as a “hit” (H). If no predicted event was made within the time window of the observed event, the event was classified as a “miss” (M). If no observed event occurred within the time window of a predicted event, the prediction was classified as a “false alarm” (FA). The sum of the hits and misses is the total number of observed events while the sum of the hits and false alarms is the total number of predicted events.

These definitions were used to compute several basic performance metrics. The “hit ratio” (HR) is defined as

$$HR = H/OE = H/\,(H + M) \qquad (1)$$

The “miss ratio” (MR) is defined as

$$MR = M/OE = M\,/\,(H + M) \qquad (2)$$

Similarly, the false alarm ratio (FAR) is defined as

$$FAR = FA/PE = FA/\,(H + FA) \qquad (3)$$

In addition to these three metrics, a “bias ratio” (BR) is defined as

$$BR = PE/OE \tag{4}$$

The hit, miss and false alarm data can be combined into a composite metric known as the critical success index (CSI; Wilks 1995). The CSI is defined as

$$CSI = H/(H + M + FA) \tag{5}$$

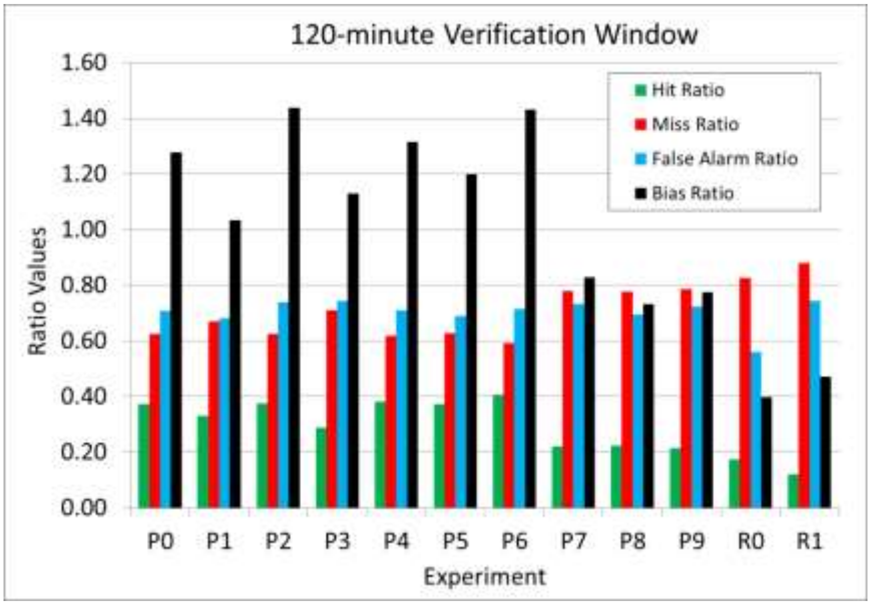
A CSI value of zero indicates that there are no hits and therefore there is no forecast skill. If the CSI is one, all the observed events were predicted with no false alarms and indicate perfect forecast skill. The CSI values are sensitive to the somewhat arbitrary choice of “hit criteria.” A forecasted event that occurs outside the specified time window will be penalized twice, first from classifying the ramp event as a missed forecast, and second from a false alarm, since the ramp event was predicted outside the ramp window. Hit, miss and false alarm events were identified for every 15-minute interval of a 15-hour forecast (for a total of 60 verification periods).

The HR, MR, FAR and BR data for the sensitivity experiments are depicted in Figure 4 and the corresponding CSI values are shown in Figure 5. Several well-defined forecast performance patterns are evident in these charts.

- Whereas the experiments with the turbine drag submodel (P7, P8 and P9) yielded the best performance on the time series metrics, the related performance on the ramp forecasts was noticeably worse than the experiments without the turbine drag submodel (P0-P6).
- The best CSI score was achieved by Experiment P5, which was the baseline configuration with the Goddard long- and short-wave radiation scheme.
- The bias ratio for the experiments with the turbine drag submodel was about 0.8 indicating they predicted substantially too few ramp events, whereas the experiments without the turbine drag submodel had a bias ratio that was generally higher than 1.0 indicating they predicted slightly too many events.
- The NWS RAP model produced ramp event prediction scores that were generally worse than the WRF configurations that did not employ the turbine drag and characterized by a bias ratio of 0.4, which indicates a severe underprediction of the number of events.

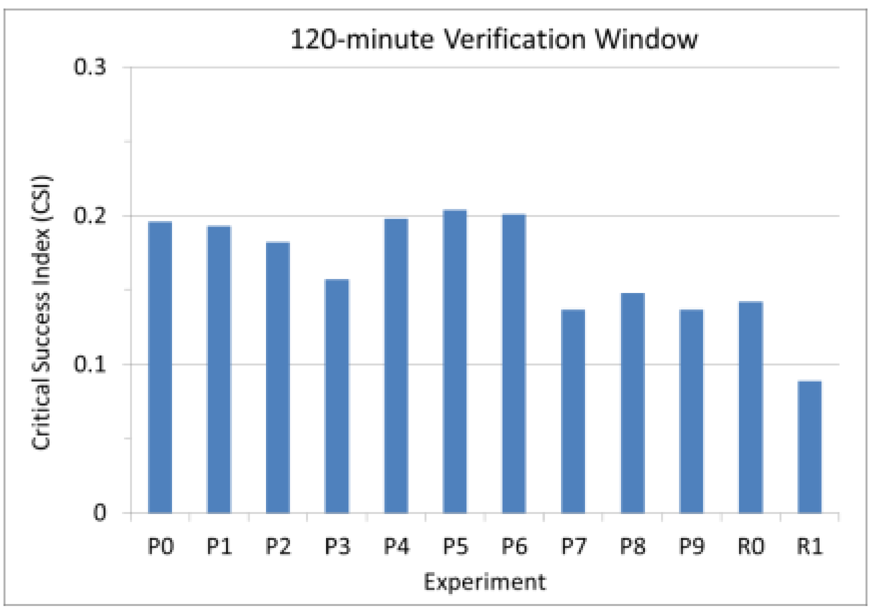
The conclusion of the performance assessment of the ramp event forecasts was 1 km WRF simulations did not outperform the NWS RAP model in the basic evaluation of the time series forecasts; however, this forecast had considerable skill over the RAP in predicting of ramp events. Furthermore, the addition of the turbine drag submodel tended to decrease the performance of the ramp event forecasts. This suggests that the outstanding performance in reducing the bias of the time series forecasts may have at least partially been a better answer for the wrong reasons.

Figure 4: Evaluation Metrics for Ramp Event Forecasts



Hit, miss, and false alarm and bias ratios for ramp events verified using a 120-minute window from Experiments P0-P9, R0, and R1. Experiment R2 is omitted because there were an insufficient number of cases completed to include in ramp verification.

Figure 5: CSI Scores for Ramp Event Forecasts



Critical success index (CSI) for ramp events verified using a 120-minute window from Experiments P0-P9, R0, and R1. Note Experiment R2 is omitted because there were an insufficient number of cases completed to include in ramp verification.

2.3.3 Vertical Profile Evaluation

The availability of the vertical wind profile data from the project sensor network provided a unique opportunity to evaluate the vertical wind profiles produced by each of the model configuration experiments. An evaluation of the bias, MAE and RMSE of the forecasted vertical wind profiles at three of the sensor sites indicated that the experiments with the turbine drag submodel produced the best forecasts in the lower layers (around 0-200 m). As might be expected, the largest difference was at the sensor location within the turbine arrays. There was a noticeable effect, however, on the verification statistics at the other two sites as well. At higher levels, the differences in model performance were significantly reduced with a great similarity in the error profiles among most of the experiments. This suggests that, as might be expected, the differences in the treatment of the physical processes near the surface of the earth are the crucial differences among the model configurations. In general, the model configurations that employed the turbine drag submodel produced the most accurate vertical profiles over all three sensor sites.

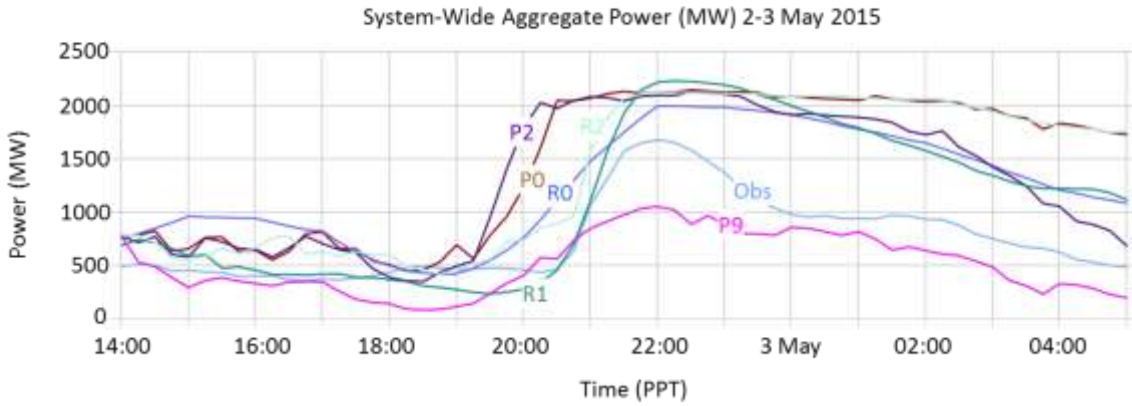
2.3.4 Case Example

The research team conducted a subjective analysis of the forecast performance for several of the ramps. This analysis gathered insight on prediction characteristics for the timing, amplitude and structure of the ramps which is often hard to infer from the quantitative statistics. The team analyzed several cases but only the May 2-3, 201, case is presented as an example in this report.

The time series of the measured and forecasted TWRA aggregated power generation for the May 2-3 case is shown in Figure 6. The most significant feature of this case is an upward ramp event that had a magnitude of 1,108.6 MW in 60 minutes and began at 2015 Pacific Daylight Time (PDT). The actual event is depicted by the light blue line labeled “Obs.” This represents the aggregated power production data from each of the wind generation facilities. In addition to time series of the measured data, forecasted time series values from the R0, R1, P0, P2 and P9 forecasts are also shown. The R1 forecast (the project version of the RAP configuration) has the best timing of the event but the associated amplitude is much larger than the measured event. The R0, P0 and P2 forecasts predict a start time that is 30 to 60 minutes before the actual start time and they have an amplitude that is too high. A signature of the event is clearly present in the P9 forecast but the amplitude is not sufficient to qualify as a prediction of a ramp event and it is the only one of the depicted forecasts that would have been classified as a “miss” of the event. For the entire forecast period, the P9 forecast has the lowest bias, MAE and RMSE.

Examining the wind profiles from remote sensing devices at three of the project sites revealed all the experimental forecasts suffered from an inability to accurately simulate the temporal progression of the event. The measured data indicated a wind acceleration event that had a temporal progression from NW to SE across the pass. However, all the NWP forecasts tended to accelerate the winds more closely to the same time at all three sites with very little evidence of a temporal progression. In addition, the NWP forecasts also tended to have erroneously high amplitude for the acceleration and wind speeds that persisted at high levels too long after the event. All of this suggests that while the forecasts of the actual event might be considered a “hit,” there is much evidence that the key processes were not accurately simulated. Ironically, the P9 simulation provided the most accurate forecast of the process but was the only forecast not to score a “hit” because the forecasted ramp amplitude was too low.

Figure 6: Time Series of Wind Generation for May 2-3, 2015, Case



Time series of aggregate power (MW) illustrating a sample diurnal up ramp along with 15-h forecasts from selected sensitivity experiments. Observed power (Obs) is plotted along with experiments P0, P2, P9, R0, R1, and R2 from 1400 Pacific Prevailing Time (PPT) May 2 through 0500 PPT May 3, 2015.

2.4 Conclusion

The team selected the configuration of the WRF model that was to serve as the baseline model for the one-year experiment to evaluate the forecast system improvements developed in this project. The selected configuration of the model and the version that was used as the baseline (i.e., the starting point) for the sensitivity experiments are listed in Appendix A. The selected configuration deviates from the widely used baseline configuration in four ways: (1) using the turbine drag submodel, (2) substituting the Pleim-Xiu Land Surface Model (LSM) for the NOAA LSM, (3) using the MYNN atmospheric boundary layer model in place of the YSU model, (4) replacing the Lin water phase change model with the WSM6 model, and (5) using the Goddard long and short-wave radiation model in place of RRTM and RRTMG models. The selected configuration produced power generation forecasts for the aggregate of TWRA wind generation facilities that had a 41.6% lower MAE and a 36.2% lower RMSE and also eliminated 88% of the bias relative to the baseline configuration over the 30-case experimental sample. It also produced a better simulation of the evolution (sequence and magnitude of the wind features at different sites) for most of the subjectively analyzed ramps.

CHAPTER 3: Field Measurements

3.1 Introduction

The main objectives of the Atmospheric Measurements and Modeling of the Tehachapi Wind Resource Area (AMMT) field campaign were to provide a robust set of meteorological data to (1) characterize the meteorological processes that influence lower boundary-layer winds in and around the Tehachapi Wind Resource Area (TWRA) and (2) improve short-term and very short-term wind forecasts (up to 15 hours and from 0 to 3 hours).

To meet these objectives, Sonoma Technology (STI) leveraged and augmented the operational California Energy Commission WindSense meteorological instrumentation network operated the instruments from May 1, 2015, through June 30, 2016 and supplied data in real time for assimilation into a meteorological model run by project partners. The main project elements included

- Making adjustments to the operations of the existing WindSense sites.
- Installing a ceilometer near the Tehachapi Airport.
- Operating all instruments, performing periodic maintenance, and performing emergency repairs, as required.
- Quality controlling all winds, temperature, and boundary layer data.
- Delivering the data set to project participants.

This chapter summarizes all activities, sites, operations, procedures followed during operation, and data completeness.

3.2 Measurements Overview

The instruments for this study were selected because they provide continuous meteorological information throughout the lower troposphere, making the resulting data appropriate for improving meteorological forecasts. Figure 7 shows the site locations for the instruments. The instruments originally installed as part of the WindSense meteorological network included:

- One Vaisala 915 megahertz (MHz) radar wind profiler (RWP) and radio acoustic sounding system (RASS) for continuously measuring profiles of winds and temperature from about 100 to 3,000+ m above ground level (AGL). The RWP and RASS were located at Bena Landfill at the southern end of the San Joaquin Valley.

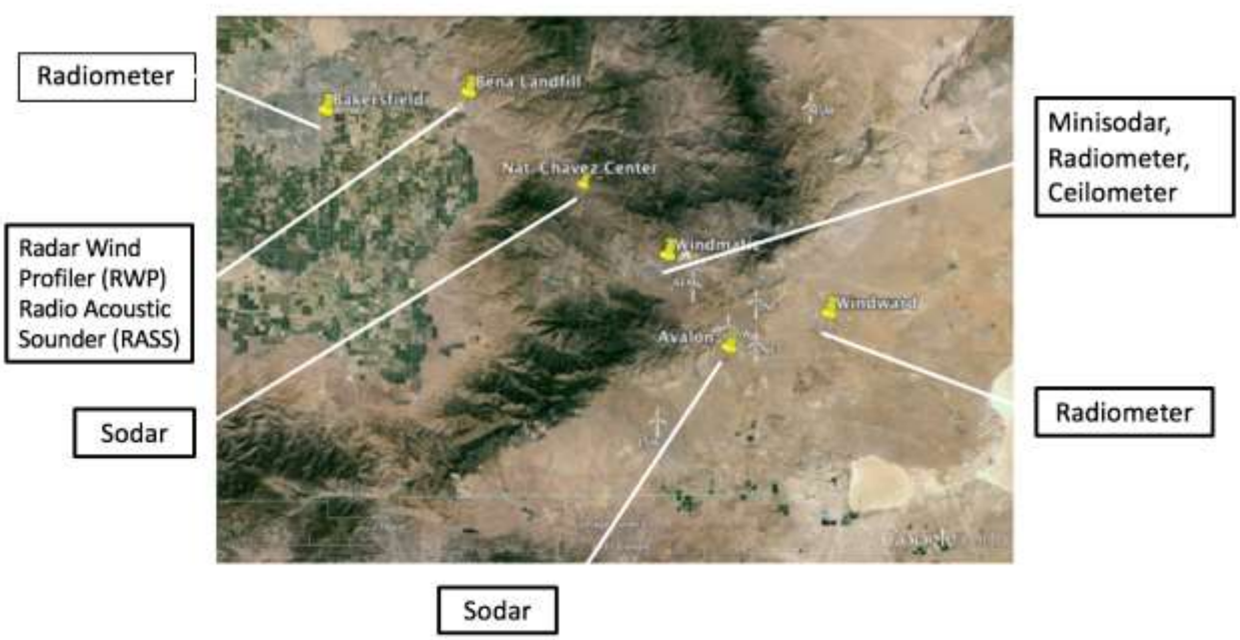
- Two Atmospheric System Corporation (ASC) Sodar 2000s for continuously measuring winds up to 600 m AGL. One Sodar 2000 was located at the National Chavez Center (Chavez) near Keene and the other was located in the Mojave Desert at a site called Avalon, which is also referred to as EDF.

- One ASC Sodar 4000 (a minisodar) for continuously measuring winds up to 200 m AGL. The minisodar was located at Windmatic near the Tehachapi Airport.

- One Radiometrics microwave radiometer for continuously measuring tropospheric profiles of temperature, humidity, and liquid water profiles. The radiometer was located at Windmatic near the Tehachapi Airport.

The research team supplemented the network by installing and operating one Vaisala CL31 ceilometer at Windmatic near the Tehachapi Airport for continuous measurements of cloud base and mixing heights. The instruments and associated measurements are summarized in Table 1, with more detailed descriptions provided in Appendix B. All the wind and temperature data were collected, archived, distributed to modelers, and displayed on a project website in real time.

Figure 7: AMMT Study Sites

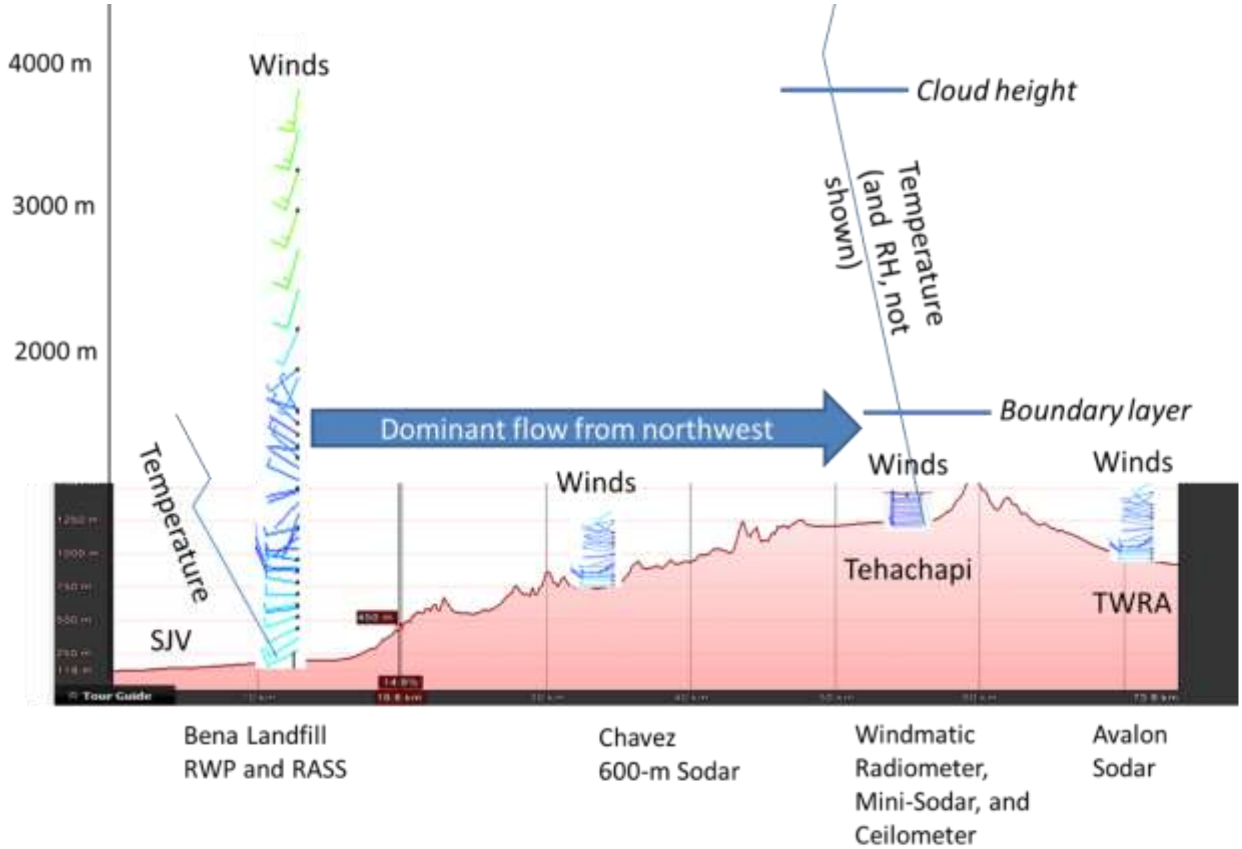


3.2.1 Rationale for Instrument and Site Selection

Meteorological conditions just upwind of the TWRA and near the top of the Tehachapi Mountains have a strong influence on conditions within the TWRA. Strong winds for power generation typically come from the San Joaquin Valley (SJV) to the northwest, up the Tehachapi Mountains (including through a corridor along California State Route [Highway] 58), and into the Mojave Desert to the south. Figure 8 illustrates this flow, the locations of instruments used to capture the vertical wind and temperature characteristics along the typical flow trajectory, and examples of the type of data obtained by the instruments measured along a typical flow trajectory.

During the warm season, winds are predominantly driven by horizontal temperature differences across Tehachapi Pass, by the stability of the air in and above the pass, and by the interaction of these factors with what is typically a slow-changing larger-scale atmospheric structure. Under these conditions, stable cold air flow can be trapped below a subsidence inversion, confining flows to passes and limiting the amount of flow over mountaintops.

Figure 8: Schematic of the Locations of Instruments and the Data Obtained



In the cold season, the large-scale progression of perturbations (i.e., storms) in the midlatitude westerly flow results in episodic events with high winds. The key forecast issues under these conditions are related to the interaction of the large-scale storm circulations with the terrain of Tehachapi Pass and nearby mountains and valleys. Phenomena such as high amplitude and breaking mountain waves can have an important effect.

To help characterize all these phenomena, one RWP was located at the southern end of the San Joaquin Valley to capture winds well upwind of the TWRA and through a deep layer (up to 4000 m AGL); one of the Sodar 2000s was located in the relatively deep canyon that follows the Highway 58 corridor, to capture winds up the main flow path for air traveling from the San Joaquin Valley over the Tehachapi Mountains; one minisodar was located near the top of the pass near Tehachapi Airport, to capture winds in the lowest 200 meters of the boundary layer just upwind of the TWRA; and one Sodar 2000 was located in the TWRA to capture winds exiting the pass and impacting the TWRA. To characterize the vertical temperature structure along the typical flow path, temperature profiles were measured using a RASS at Bena Landfill and microwave radiometers near the Tehachapi Airport and the Bakersfield Airport, and at the Southern California Edison Goldtown Substation in the Mojave Desert. The microwave radiometer also provided vertical profiles of humidity and liquid water. The diurnal and spatial pattern of boundary-layer heights and entrainment of aloft momentum can have a dramatic impact on winds in the TWRA. As part of this project, a ceilometer was installed at the

Windmatic site to continuously measure cloud base and mixing-layer height. Wind profiler, sodar, and ceilometer data were postprocessed to determine continuous boundary-layer heights and structure (Table 1).

Table 1: Measurements Collected During the AMMT Field Study

Instrument / Manufacturer	Parameter(s)	Site Name	Location (degrees) and Elevation (m MSL)	Measurement Height(s) Above Ground Level (m)	Vertical Resolution (m)	Frequency (min)
Mini-Sodar (ASC Sodar 4000)	Wind speed and direction	Windmatic	35.12707, -118.4263; 1231	~20 to 200	10	15
Sodar (ASC Sodar 2000)	Wind speed and direction	Avalon (EDF)	35.00205, -118.32473; 1049	~80 to 600	25	15
Sodar (ASC Sodar 2000)	Wind speed and direction	National Chavez Center	35.22784, -118.56413; 782	~80 to 600	25	15
Microwave Radiometer (Radiometrics MP-3000A)	Temperature, humidity, and liquid water	Windmatic	35.12707, -118.4263; 1231	~10 to 10,000	50 below 500 100 between 500 and 2000 250 above 2000	6
RWP (Vaisala LAP 3000)	Wind speed and direction	Bena Landfill	35.34962, -118.75807; 345	~120 to 3500	60 to 100	55
RASS (Vaisala LAP 3000)	Virtual Temperature	Bena Landfill	35.34962, -118.75807; 345	~120 to 1500	60	5 at top of hour
Ceilometer (Vaisala CL31)	Boundary layer and cloud base height	Windmatic	35.12707, -118.4263; 1231	~2 to 10,000	10	1

3.3 Instrument Preparation, Installation, and Operations

The objective of routine instrument operations is to ensure high-quality data and high data recovery rates. Operations for this project were divided into two main elements: predeployment instrument interface and testing, and routine operations.

The RWP/RASS, sodars, radiometer, and minisodar were already operational at the start of this project. In this final phase of the project, the only new instrument deployed was the ceilometer. To prepare the ceilometer for deployment, the power source, computer, data management system, and communications were tested at STI and system corrections were made as needed to ensure that the ceilometer met manufacturer specifications and that all systems, from data collection to data delivery and archiving, worked together properly. STI's team installed the ceilometer and supporting system at the field site in accordance with manufacturers' guidelines and field tested the equipment to ensure that all components, including the power source, computer, data management system, and communications, worked properly.

Maintenance of all instruments is critical to successful operations. STI's team conducted routine maintenance and made emergency site visits when necessary. Routinely, meteorologists at STI's Weather Operations Center compared the surface and upper-air data from each site with external data sources as a quality control measure, allowing for the identification of any operational or equipment problems.

3.4 Data Flow and Processing

Reliable communications with each site are required to ensure high data recovery rates for real-time data use, to monitor instrument performance, to remotely diagnose instrument problems, and to make instrument system changes as needed. This goal was achieved by using cellular communications and file transfer protocol (FTP) at each site.

Dual-band cellular modems were located at each site and STI automatically pushed data every 30 minutes (60 minutes for the RWP/RASS) from each site to the FTP servers. Once the data were uploaded, an automatic process took the raw data and stored them in a Microsoft® SQL Server® database, effectively combining all data into a single data set. Raw data files were stored and backed up each day. Another automatic process generated images of the data and uploaded them to the project website. At the same time, the data were provided every 30 minutes (60 minutes for the RWP/RASS), a few minutes after collection, to an FTP site for download and assimilation into the meteorological forecast model.

3.5 Data Completeness

Overall, the instruments were operated with great success and provided a very robust set of data that met the project measurement objectives.

Data completeness is defined as the percentage of time with data divided by the total number of records possible. The number of records possible was determined using the instrument installation date, operations end date, and the frequency of measurement (every 30 minutes for all instruments, except the RWP/RASS, which was 30 minutes through January 27, 2015, and hourly thereafter). Data completeness ranged from 81.5% for the minisodar to nearly 100% for the Sodar 2000 at Chavez.

CHAPTER 4: Short-Term Wind Ramp Forecasting Improvements

4.1 Introduction

The work completed in Task 4 improved low-level wind forecasts, in particular at the wind turbine hub height (80 meters), for the Tehachapi Wind Resource Area by assimilating data collected from four project sites—one wind profiler at Bena, two sodars at Chavez and EDF, and one minisodar at Windmatic—and two radiosonde (sonde) field campaigns. Analyses with different data assimilation techniques and strategies were produced and used to initialize model forecasts. The forecasted wind results were verified at the 80-meter height, primarily at the EDF site, which is within a wind plant, and in some cases at Windmatic, upstream of EDF. Ramp event forecasts were evaluated by converting observed and forecasted 80-meter wind speeds at the EDF site to electric power using the Vestas V-90 3MW wind turbine power curve. The Gridpoint Statistical Interpolation (GSI) analysis system and the WRF model were used for all experiments.

Three studies are included in this report: a one-month study (June 2016), a case-driven study (10 ramp events or nine cases that occurred during April-June 2015), and a sonde field campaign study. On a seasonal timescale, the highest number of ramps occurs in the spring, when the weather is dominated by the diurnal cycle. This study focused on springtime cases.

4.2 Methods

The primary numerical tools used in this task are the GSI analysis system and the WRF model. The GSI analysis system was used to assimilate observations, while the WRF model was used to produce weather forecasts..

4.2.1 GSI Analysis System

The GSI is a unified data assimilation (DA) system that can be used for both global and regional applications (Wu 2005; De Pondeca et al. 2007; Kleist et al. 2009). The GSI is a joint effort among many collaborators, though it has been developed mainly by scientists at the National Centers for Environmental Prediction (NCEP). The system can produce analyses using different data assimilation methods, such as the 3D-Var method, an ensemble square root filter (EnSRF), and a coupled EnSRF–three-dimensional ensemble-variational hybrid (EnSRF–En3DVar) method (Pan et al. 2014). In this project, all three methods were tested and compared. The GSI can work with different numerical forecast models and has the flexibility to incorporate new developments such as new observational types, improved data quality control, new analysis variables, anisotropic background error covariances and expansion to four-dimensional variational data assimilation. The GSI v3.3 for 3D-Var and hybrid and v3.5 for EnSRF were used in this task.

4.2.2 Experiment Design

The physics schemes that were used in the WRF forecasts are described in Appendix C. The WRF model land use data were updated with the European Space Agency 2010 global land cover dataset, which has a spatial resolution of 300 m. All the numerical experiments conducted in this task use the new land use data. Due to limitations on computational resources, only one domain with a resolution of 3 km was used for DA and forecasts. The horizontal spacing was 3 km, and the vertical spacing was stretched with a higher resolution in the lower atmosphere. The domain covered Southern California, centered at the TWRA. Three sets of observations were used: NCEP Global Data Assimilation System (GDAS) data available in 6-hour intervals, project wind data available in 1-hour intervals, and radiosondes launched at roughly 3-hour intervals from the two sonde field campaigns.

While several numerical studies were conducted during the project, only three of them are reported here: a) one month of analyses and 12-hour forecasts for June 2016, b) 10 ramp events (9 cases) listed in Appendix C and c) two high wind events from the sonde field campaigns.

4.2.2.1 June 2016 Study

The team first evaluated data assimilation (DA) techniques and strategies using a one-month (June 2016) study. Two observational datasets were used: GDAS and project winds from the four project sites over the TWRA. The NCEP Rapid Refresh (RAP) data were used to provide initial and boundary conditions. Data assimilation cycles were performed from 0000 Coordinated Universal Time (UTC) May 30, 2016 to 1200 UTC June 29, 2016. Twelve-hour forecasts, which were initialized by DA analyses, were produced every six hours, starting from 0000 UTC June 1 until 1200 UTC June 29, 2016. The DA cycling procedure is described in Appendix C. During the data assimilation period, GDAS data were assimilated every six hours, while the project data were assimilated at intervals depending on the experiment design.

4.2.2.2 Ten Ramp Event Study

This study examined the 10 diurnal ramp event cases from April-June 2015 that were identified in Task 2. One pair of upward and downward ramps occurred on the same day, so the 10 events were studied as nine forecast cases. In this event-driven study, the forecast length was 24 hours, instead of 12 hours as used in the June 2016 study. Two sets of observations—GDAS and project data—were assimilated. For each case, four sets of numerical experiments were conducted that differed either in the data assimilation technique (i.e., the 3DVAR, ENKF, and HYBRID experiments) or assimilated data. The set of NO_PROJ experiments assimilated only GDAS using the GSI 3D-Var technique. DA cycling was conducted as described in Appendix C.

4.2.2.3 Radiosonde Field Experiments

Two sonde field campaigns were conducted during high wind events in the summer of 2016. The first campaign took place from 0000 UTC June 25 to 1800 UTC June 25. Radiosondes were launched at two sites: one in Bakersfield and the other in Tehachapi. The second campaign took place from 1600 UTC July 23 to 1800 UTC July 24. Radiosondes were launched at two sites: the same Tehachapi site as the first campaign, and the EDF site. Radiosondes were launched approximately every three hours. Sounding data were collected until the weather balloons exploded at high altitude, approximately 100 hPa.

Nine data assimilation experiments were conducted for the radiosonde cases, using three DA methods (3D-Var, EnSRF and hybrid) and three groups of observational data. The basic set of observations included conventional sources such as land surface stations, buoys, and ships, and satellite radiance data. The next set of observations added was from project sensors, while the third set came from the radiosonde launches. A summary of the nine DA experiments is given in Appendix C. For each experiment, a 24-hour forecast was made after six hours of DA cycling.

4.3 Results

4.3.1 Data Assimilation Analysis

To verify whether the observations had been assimilated properly, statistical comparisons were conducted between O – B (observation minus background) and O – A (observation minus analysis). Successful data assimilation should reduce O – A, which incorporates the assimilated data, compared with O – B, which does not.

4.3.1.1 June 2016 Study

Figure 9 shows the time-height cross-sections of wind analyses at the EDF site for one month. The observations clearly present a diurnal pattern; however, the daily extremes of wind speed vary significantly. All the DA analyses show similar diurnal patterns to the observations. The experiments that assimilated project wind data were able to reproduce the observed wind magnitude better than the one without (i.e., NO_PROJ), which was expected since observations from the EDF site were assimilated. Due to the lack of observations from June 23-25, all DA analyses show very strong wind speeds, which are very likely overestimated.

Table 2 shows the one-month statistics of O-B and O-A using different data assimilation strategies. The data at the wind profiler site (i.e., the Bena site) were not used because there were too many missing observations. The root mean square (RMS) values of O-A are smaller than those of O-B for each experiment. This result implies that observations were properly assimilated in the GSI analysis system. Among the four experiments, 3DVAR_3H has the smallest values for both O-A and O-B for this specific study month.

Table 2: RMS of O-B and O-A for June 2016 With Different DA Methods

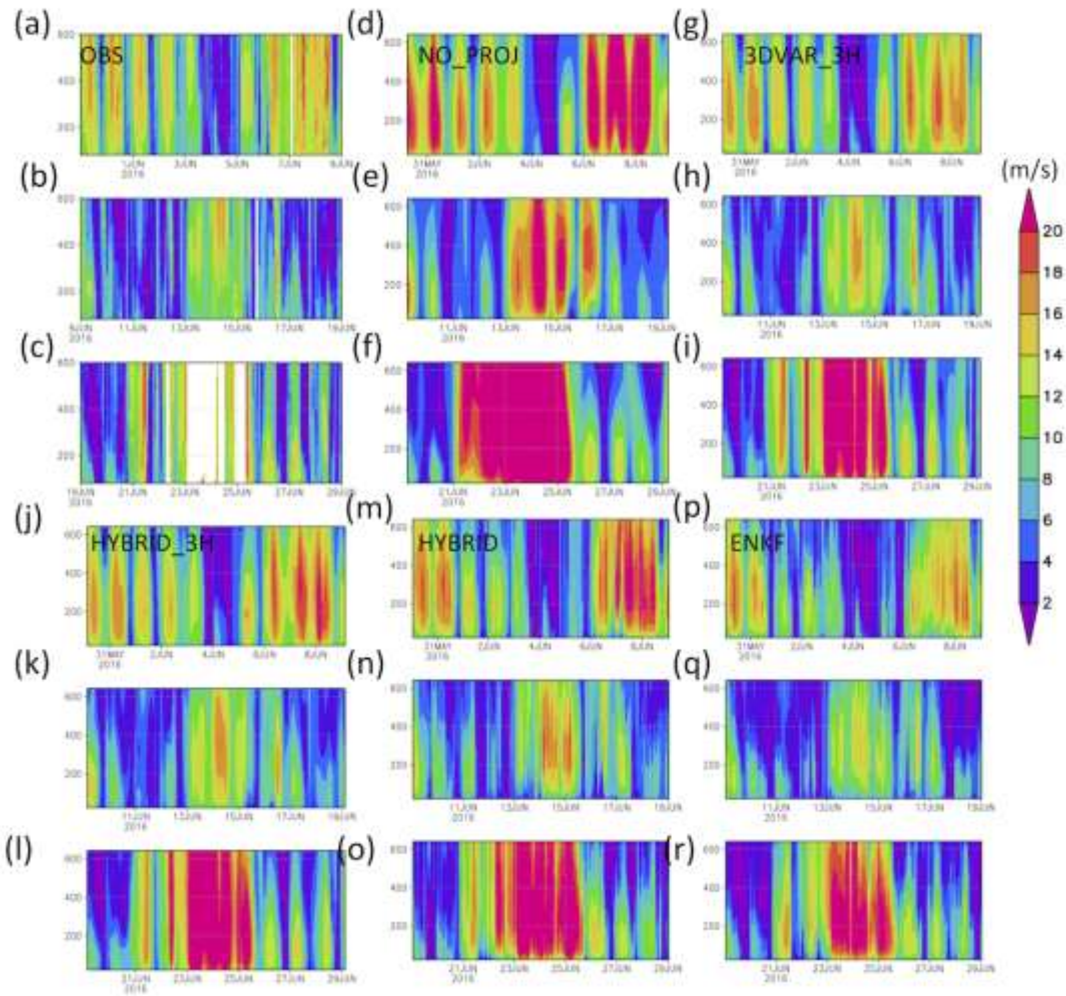
	NO_PROJ	3DVAR_3H	ENKF	HYBRID_3H	HYBRID
O-B		2.34	3.36	2.36	2.51
O-A		0.90	2.54	1.38	1.80

Unit: m s^{-1}. Note that results from the Bena site were excluded because too many data were missing.

4.3.1.2 Ramp Event Cases

Table 3 presents the RMS statistics of O-B and O-A from the nine cases using points at project sites. The RMS values of O-A are smaller than those of O-B and there is a comparable amount of data in the three experiments. Experiments with the assimilation of project data (i.e., 3DVAR and HYBRID) gave analyses closer to observations than without at the EDF site.

Figure 9: Time-Height Cross Sections of Wind Speeds (shaded; m/s) at the EDF Site



From 0000 UTC 30 May 2016 to 0000 UTC 29 June 2016 at the EDF site. (a-c) Sodar observations, (d-f) NO_PROJ analysis, (g-i) 3DVAR_3H analysis, (j-l) HYBRID_3H analysis, (m-o) HYBRID analysis, and (p-r) ENKF analysis

Table 3: RMS of O-B and O-A for the Ramp Event Cases With Different DA Methods

	NO_PROJ	3DVAR	ENKF	HYBRID
O-B		3.37	3.61	3.95
O-A		1.35	1.70	1.42

Unit: m s^{-1}. Note that results from the wind profiler (Bena) site were excluded because too many data were missing.

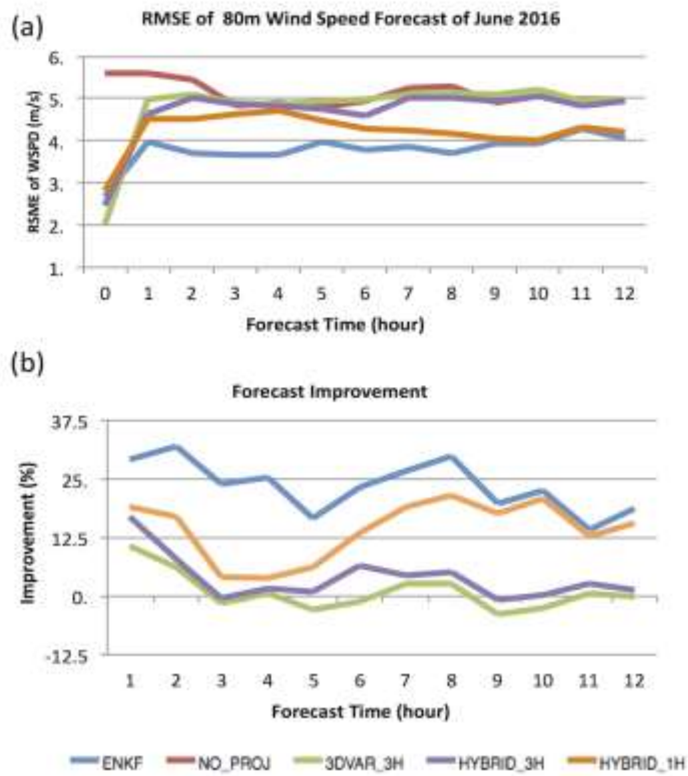
4.3.2 Forecast Performance

4.3.2.1 June 2016 Study

The experiment without project data (NO_PROJ) often over-forecasted wind speed, in particular when high winds were observed. The root mean square errors (RMSEs) of the 80-m wind speed forecasts at the EDF site (Figure 10 and Table 4) were slightly lower for 3DVAR_3H and HYBRID_3H than for NO_PROJ. The reduction can be attributed mainly to improvement in the first two hours. The RMSE of 3DVAR_3H (HYBRID_3H) forecasts was reduced by 4% (8%) in the first hour and 7% (8%) in the second hour. The brief duration of the improvement was expected since the project observations are confined to a limited region near the wind plant areas.

The assimilation of project data at a higher frequency, i.e., hourly, clearly improved high-wind forecasts at the 80-m height, though high winds were sometimes still overforecasted, in particular when many project data were missing. The RMSE of the wind speed at 80 m was reduced throughout the entire 12-hour forecast period by an average of about 14% for the HYBRID experiment and 24% for the ENKF experiment.

Figure 10: Error Statistics of 12-h, 80-m Wind Forecasts at the EDF Site - June 2016



(a) RMSE and (b) forecast improvement with respect to the NO_PROJ experiment

Compared to the EDF site, the statistics of 80-m wind forecasts at the Windmatic site are different. The Windmatic site is located upstream of the EDF site and has little to no influence from wind turbines. The ENKF experiment produced the largest RMSE of 80-m wind forecasts, with a negative bias. The negative wind bias in ENKF is because the initial conditions were taken from the ensemble analysis mean, which was smoother and weaker compared to the other experiments. Thus, the wind forecasts from ENKF were weaker than others at the Windmatic site. At the EDF site, the winds were underestimated because the smoother and weaker initial conditions were compensated for by overestimating the wind from the neglect of the wind turbine effect in the model.

Table 4: Error Statistics of 80-m Wind Forecasts for the June 2016 Experiments

(a)

	NO_PROJ	3DVAR_3H	ENKF	HYBRID_3H	HYBRID
MAE	3.85	3.78	3.12	3.73	3.51
RMSE	5.08	5.02	3.88	4.87	4.34
BIAS	2.10	1.94	-0.79	2.10	0.00

(b)

	NO_PROJ	3DVAR_3H	ENKF	HYBRID_3H	HYBRID
MAE	1.978	1.785	3.014	1.728	2.078
RMSE	2.382	2.182	3.578	2.068	2.528
BIAS	-1.113	-0.982	-2.622	-0.899	-1.522

(a) EDF site and (b) Windmatic site. MAE is the mean absolute error; RMSE is the root mean square error. Unit: m/s

4.3.2.2 Ramp Event Cases

For each of the nine cases, four experiments were conducted: NO_PROJ, 3DVAR, ENKF and HYBRID. Among the four experiments, NO_PROJ and 3DVAR produced very similar error patterns at the EDF site. The 3DVAR performed slightly better than NO_PROJ; the errors in 3DVAR were slightly smaller for the first hour at 80-m and for the first five hours at 100-450 m. ENKF and HYBRID had more pronounced improvements in the first five forecast hours. After about eight hours of forecasts, however, the errors between 200 to 450 m became about 1.5 – 2 times larger than those from NO_PROJ and 3DVAR, and the errors became larger with increasing elevation. Further investigation shows that the negative wind bias at upper levels in ENKF and HYBRID was the main cause of the high RMSE. ENKF produced the best 80-m wind forecasts at the EDF site because the initial conditions (i.e., ensemble analysis mean) were smoother and weaker, which compensated for the absence of the effect of wind turbines in the model.

The MAE and RMSE at the Windmatic site are smaller than those at the EDF site because the observed winds were weaker. Similar to the June 2016 experiments, ENKF and HYBRID gave the largest errors (~ 2-3 m s^{-1}) because of the related negative wind biases. Both 3DVAR and NO_PROJ had almost zero bias for 80-m wind forecasts at the Windmatic site.

4.3.3 Ramp Event Detection

4.3.3.1 Ramp Event Detection

Ramp events are defined based on changes in power production. Thus, the observed and forecasted wind speeds at 80-m height at the EDF site are converted into estimates of power output using the Vestas V-90 3MW wind turbine power curve with the V-90 blade radius (Fitch, AC et al. 2012). The team used the wind ramp definition in Zack et al. (2010) to detect events satisfying the following two criteria. First, the power change exceeds 30% of the power at the previous hour. Second, the time of the forecasted ramp event is allowed to shift one hour before or after the actual time of the occurrence.

4.3.3.2 Ramp Event Forecasts for June 2016

Because of underpredicting wind speed in ENKF, which can bias the scores, it is excluded in this section. Table 5 shows the scores of forecasted ramp events from different numerical experiments for the June 2016 study. All biases are greater than 1, indicating that the model tended to overforecast the ramps. The HR and CSI from 3DVAR_3H are twice as high as those from NO_PROJ, indicating the benefit of assimilating project data for ramp event forecasts. The forecast scores from HYBRID_3H are better than those from 3DVAR_3H and NO_PROJ. The bias is significantly reduced using the hybrid method. Among all the experiments, HYBRID_3H produced the best ramp event forecasts, including the highest HR, the lowest FAR, the lowest bias and the highest CSI.

Table 5: Scores of Ramp Event Forecasts - Numerical Experiments, June 2016 Study

	HR	FAR	Bias	CSI
NO_PROJ	0.035	0.965	2.714	0.026
3DVAR_3H	0.089	0.911	2.500	0.068
HYBRID_3H	0.176	0.824	1.214	0.107
HYBRID	0.129	0.871	1.518	0.085

4.3.3.3 Ramp Event Forecasts for Selected Cases in 2015

Table 6 shows the scores of forecasted ramp events from different numerical experiments for the 10 selected ramp events in 2015. Again, ENKF experiments are excluded in this evaluation. The conclusions are similar to those in the June 2016 study. Both 3DVAR and NO_PROJ produced high biases (> 4), indicating significant overforecasts of the ramp events. NO_PROJ has the highest FAR (more than 90%). HYBRID has an HR of more than 20%, which is higher than HYBRID in the June 2016 study and is the best forecast among all the experiments.

Table 6: Scores of Ramp Event Forecasts From Numerical Experiments - Nine Cases (2015)

	HR	FAR	Bias	CSI
3DVAR	0.132	0.868	4.222	0.119
HYBRID	0.209	0.791	1.899	0.158
NO_PROJ	0.099	0.900	4.782	0.089

4.3.4 Field Case Study

In the June case, the observed wind speed decreased by about 12 m s^{-1} during the first 12-hour period, and it increased again by roughly 8 m s^{-1} during the next 12-hour period. This overall variation was well simulated in all nine DA experiments. However, the initial wind speed in all the DA experiments is larger than the observation by about 8 m s^{-1}, and there is a 3-4 hour delay before the wind speed starts increasing. The timing of the wind speed increases in EXP6 (EnSRF with all observation types) and EXP9 (hybrid with all observation types) are one-hour closer than the other experiments.

In the July case, the observed wind speed increased during the first five-hour period, maintained its magnitude for the next 11 hours, and decreased to a minimum of 2 m s^{-1} over the next five hours. Among the three 3D-Var experiments, the wind speed error is the smallest in EXP3 (with all observation types). The simulated time series of wind speed in the EnSRF experiments is closer to the observation than that in the 3D-Var experiments. Characteristics of the time evolution of the hybrid experiments are similar to those of the 3D-Var experiments. Among the hybrid experiments, simulated wind speed in EXP9 (with all observation types) is closer to the observation than the other hybrid experiments.

In summary, when observations are assimilated using more advanced DA methods, such as the hybrid method, the wind speed forecast is improved, and this improvement is associated with the assimilation of localized special observations.

4.4 Conclusions

The major conclusions obtained from the one-month and ramp event studies are listed:

- The GSI system was able to assimilate observations reasonably well, as the RMS of O-A is smaller than the RMS of O-B. Among the three DA methods, 3D-Var and hybrid produced comparable RMSs of O-A, while EnSRF produced a larger RMS of O-A.
- The use of 3D-Var and hybrid methods with three-hour data cycles gave comparable results at the EDF site. The additional assimilation of wind data from four project sites improved the 80-m wind forecasts for the first two hours at the EDF site for both experiments. More frequent (hourly) assimilation of observations improved forecasts at the wind turbine site (EDF) but degraded forecasts upstream.
- The EnSRF method performed well at the EDF site yet had the largest errors of any method at the Windmatic site. This is because two effects compensated for each other at the EDF site, which is located within a wind plant. First, the using the analysis ensemble

mean in the EnSRF method causes underforecasts of wind speed intensity. Second, neglecting the wind turbine effect in the WRF model causes overforecasts of high wind speeds at the EDF site.

- For the ramp event forecasts, the hybrid method gave the best result, including the highest hit rate, the lowest false alarm ratio, the lowest bias, and the highest critical success index.

CHAPTER 5: Very Short-Term Wind Ramp Forecasting Improvements

5.1 Introduction

The primary objective of the development of a statistical very short-term forecast method for the project was to create a capability to generate rapid update predictions (15-minute update frequency) of the power generation time series and potential for wind ramp events for the 3-hour period following the issue time of the forecast.

This capability is motivated by the other primary tool for short-range prediction, physics-based Numerical Weather Prediction (NWP). It typically does not provide much value for very short-term look-ahead periods and is not practically suited to produce very frequent (sub-hourly) forecast updates.

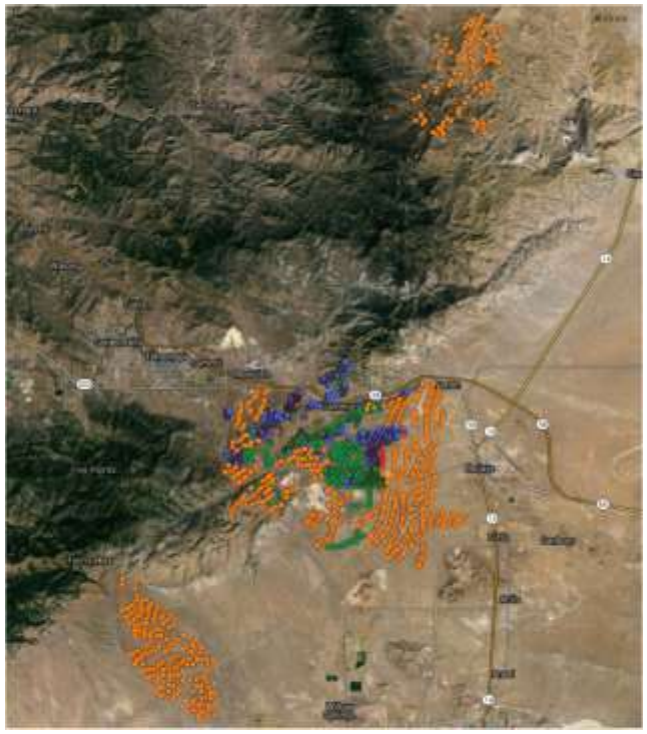
The fundamental approach used to develop the rapid update very short-term prediction capability was the use of an advanced machine learning algorithm to develop a statistical prediction model by training the algorithm with about two years of time series data from sensors in or near the TWRA.

5.2 Input Data

The sources of input data for the very short-term statistical prediction tool can be divided into three categories.

- **Wind generation facilities:** Data from 17 generation resources were used in developing and evaluating the prediction tool. The locations of the wind turbines associated with these 17 facilities are shown in Figure 11. The aggregated generation capacity of these facilities is 2319 MW. This capacity represents about 2/3 of the total wind generation capacity in the TWRA. Data from the other generation facilities were not used because a sufficient quantity of data was either not available or the quality of the data did not meet basic standards.
- **Local area nonproject data sources:** These stations are a diverse set of sensor installations operated by a wide variety of entities ranging from public agencies to individual residents. Specifications of the data obtained from each sensor are provided in Appendix D.
- **Project-deployed targeted sensor network:** These sensors are described in Chapter 3.

Figure 11: Turbine Locations Associated With the Wind Generation Facilities Used in the Short-Term Forecasting Experiments



The color of the markers indicates the turbine hub height.

5.3 Machine Learning Configuration

5.3.1 Machine Learning Method

Two machine learning methods were employed for the very short-term forecasting application: the Gradient Boosting Machine (GBM) (Friedman, 1999) and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016).

Gradient boosting builds a set of weak predictive models, one at a time. The first model predicts only the gross features of the data. Each successive model, when added to a linear combination of the prior models, adds the capability to predict successively more subtle features of the data. XGBoost is very similar to GBM but uses a more regularized model to help prevent overfitting and is able to leverage hardware to reduce run times. The version of GBM used in this project's experiments is the implemented in the Python module library in the "Scikit-learn" package (http://scikit-learn.org/stable/). The version of XGBoost used in the experiments is Version 0.6a2, currently available at http://xgboost.readthedocs.io/en/latest/. The values of the XGBoost parameters used in the optimal configuration are listed in Appendix D.

5.3.2 Predictands

The prediction tool was formulated to operate in two modes: (1) time series and (2) ramp rate. In the time series mode, forecasts of the 15-minute average power generation were produced every 15 minutes for the 0-3 hour look-ahead period. In the ramp rate mode, predictions of the maximum upward (RAMPMAX) and downward (RAMPMIN) 60-minute ramp rates expected within the next three hours were generated.

Computing the critical success index (CSI) requires setting thresholds for RAMPMAX and RAMPMIN to allow for a "yes" or "no" prediction for each threshold. Thresholds were set to 300, 500 and 800 MW for RAMPMAX and -300, -500 and -750 MW for RAMPMIN. The threshold of -750 MW as opposed to -800 MW was selected for RAMPMIN to allow sufficient sample size of observed ramps.

5.3.3 Predictors

A pool of 116 potential predictors was defined using the data available from the full collection of input sources. The predictors were formulated to include a broad array of variables with potential predictive value based on knowledge of the meteorological processes that drive ramps in the TWRA. These include

- observed power production from 0 to 90 minutes before the forecast issue time.
- the vertical wind shear and vertical temperature gradients in and just upstream of the TWRA, 3) the wind component from 300° (i.e., through the pass) from the surface to the 700 millibars (mb) pressure level in and upstream of the TWRA.
- the presence of wind speeds approaching or exceeding the cut-out threshold for wind turbines in the TWRA.
- the pressure gradient across the pass.
- the flow of marine air over the coastal ranges into the Central Valley.

The most recent observed values and the time rates-of-change over 15 to 120 minutes are considered. The list of all the candidate predictors and the sources of data used to compute them is provided in Appendix E.

To maximize the size of the training sample but also provide an independent sample for the evaluating the resulting forecast models, a 24-month rolling training sample approach was employed. The evaluation period was October 2015 to September 2016 (the overall target evaluation period for the project). Forecasts were generated for each month in the evaluation period by using a training sample of 24 of the available 25 months, with the month being forecasted excluded from the training sample.

A procedure was developed to select the predictors with significant predictive power from the initial pool of 116 candidate predictors. The base set of four predictors consisted of the three time variables (year, day of year and time of day) plus the most recent observed value of the predictand. This left 112 predictors available for screening process. In the first screening procedure, 112 trainings and forecasts were created, each adding only one of the unused predictors to the base set of four predictors. The single predictor with the greatest reduction in forecast MAE over the entire one-year forecast period was then added to the set of four predictors. The screening proceeded by adding each of the 111 remaining predictors to the set of five predictors already in use. The screening continued until the predictor that yielded the greatest forecast improvement resulted in less than a 0.4% MAE reduction. Separate screenings were performed for the time-series and ramp rate mode forecasts.

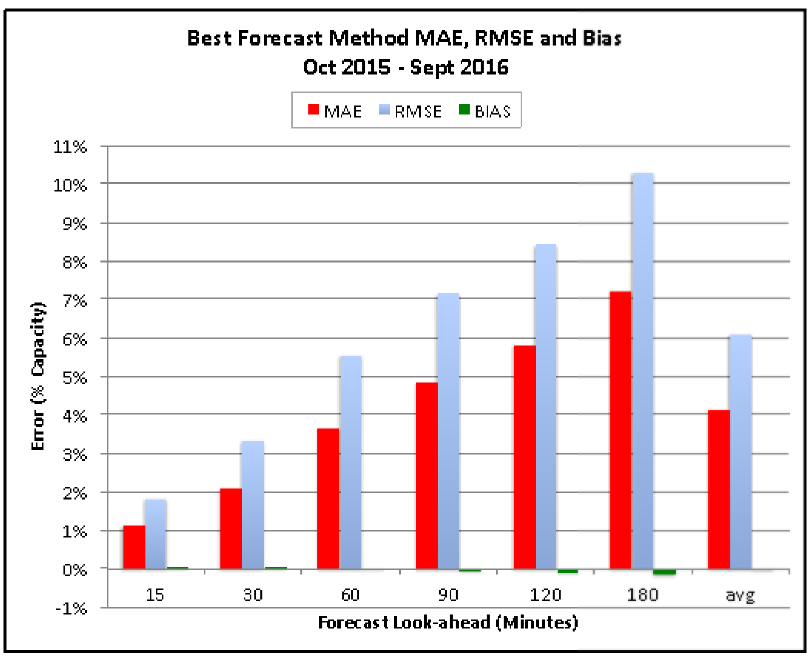
5.4 Performance Analysis: Power Generation Time Series

5.4.1 Performance of Best and Final Configuration

The best and final configuration for the time series mode of the very short-term forecast system was based on the results from the extensive number of experiments conducted. This was the configuration that was ultimately employed as one of the forecast method ensemble members in the Improved Operational Forecast System (IOFS), which represented the integrated set of forecast system improvements developed in this project. The predictors that were selected for the best and final configuration are listed in Appendix D.

The MAE, RMSE and bias for the best and final forecast system configuration over the one-year forecast evaluation period are shown in Figure 12. The MAE ranges from just over 1% of capacity for a look-ahead period of 15 minutes to just over 7% of capacity for a look-ahead period of 180 minutes. The average over the entire 0-3 hour forecast period is about 4% of capacity. The RMSE ranges from just under 2% of capacity at 15 minutes to just over 10% of capacity at 180 minutes. The average RMSE over the 0-3 hour forecast period was about 6%. Bias was less than 0.2% of capacity for all 15-minute forecast intervals in the 0-3 hour forecast look-ahead window.

Figure 12: MAE, RMSE and Bias by Look-Ahead Time for the Very Short-Term Forecast Method

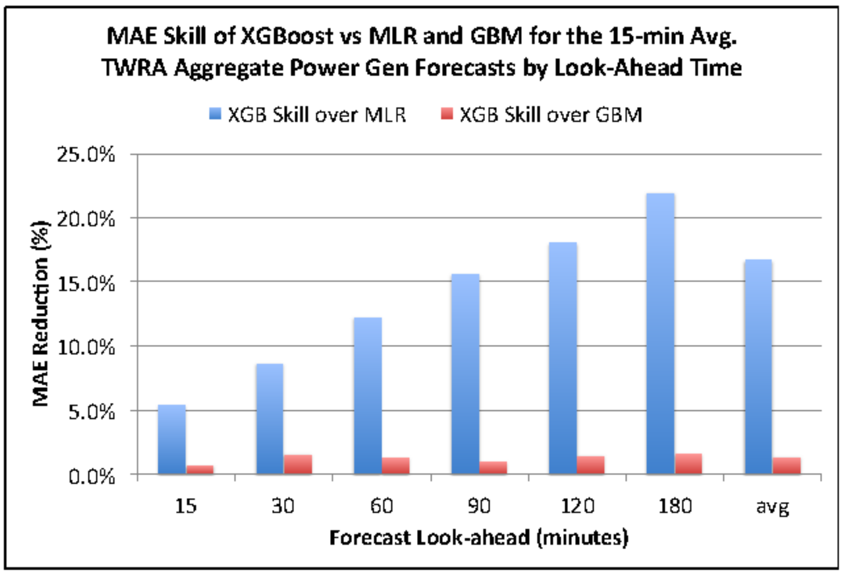


These produced the best forecasts over the 0-3 hour look-ahead period.

5.4.2 Sensitivity to Machine Learning Method

An experiment was conducted to evaluate the relative performance of XGBoost and GBM for the Tehachapi 0-3 hour wind power prediction application. In both methods, the internal parameters were optimized for this application. Figure 13 indicates that XGBoost provided on average a 3% reduction in MAE over GBM. The maximum benefit was 3% to 3.5% for the 60-120 minute look-ahead period and the minimum benefit was 1.4% for the 15 minute-ahead forecast. Based on these results, the XGBoost method was used for all subsequent 0-3 hour machine-learning-based forecast experiments in this project. Figure 13 also shows the performance of XGBoost relative to multiple linear regression (MLR). XGBoost significantly outperformed MLR with a benefit ranging from 5.4% for the 15-minute ahead forecast to 21.9% for the 180-minute ahead forecast. The average benefit was 16.7%.

Figure 13: Comparison of XGBoost to GBM and Multiple Linear Regression



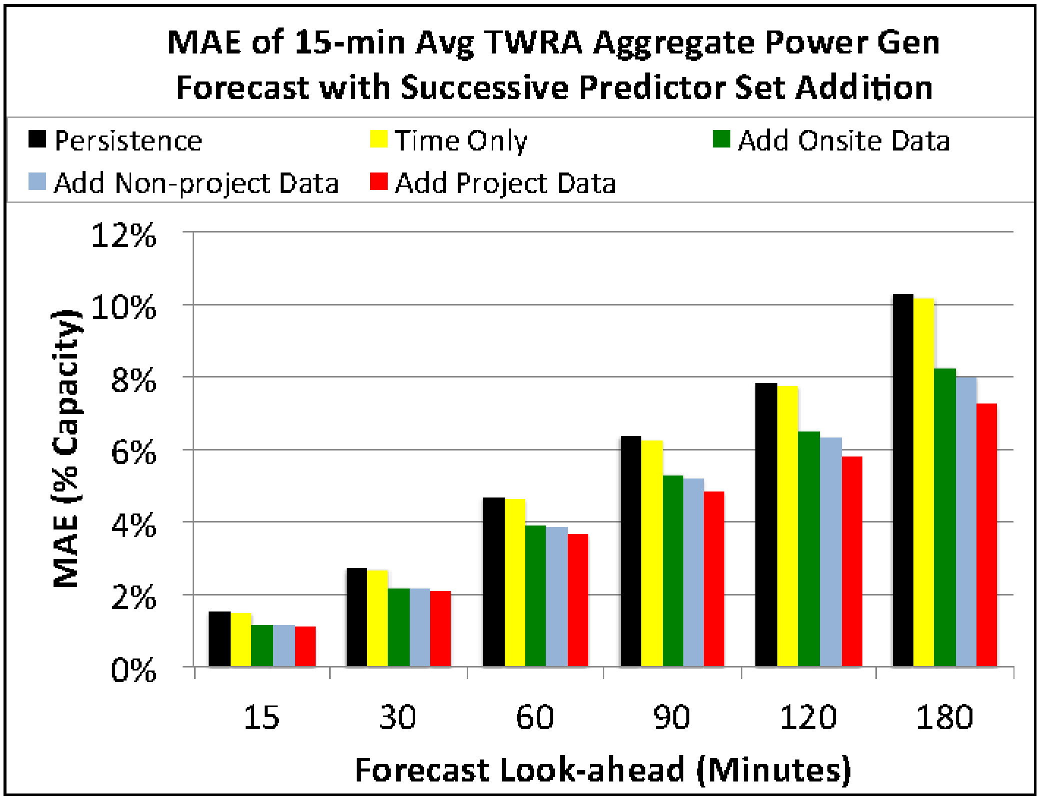
Percentage reduction in MAE when the XGBoost method is used in place of multiple linear regression and the GBM method for the production of 0-3 hour ahead forecasts of the 15-minuite average power generation by the TWRA wind generation aggregate (2319 MW) for the one-year period extending from October 2015 to September 2016.

5.4.3 Effect of Predictors by Source Category

A key objective of the forecast performance evaluation was the assessment of the forecast performance benefit obtained from the targeted network of sensors deployed in this project. This benefit was evaluated by training XGBoost models and generating forecasts from five subsets of the predictors selected by screening. The subsets were based on the source of the data used to compute each predictor. This experiment also includes a comparison of the performance of the XGBoost to multiple linear regression.

The MAE in percentage of capacity by look-ahead time for each of the predictor source subset experiments is depicted in Figure 14. The largest benefit occurs when onsite data are added to the forecast. However, non-project and, especially project data have a significantly affect at look-ahead times of 60 minutes or longer. The average MAE reduction obtained by using the project sensor data over the entire 0-3 hour forecast period is nearly 7.5%. It varies from about 2% at 15 minutes to nearly 10% at 180 minutes.

Figure 14: MAE of Power Production Forecasts by Look-Ahead Time and Predictor Source



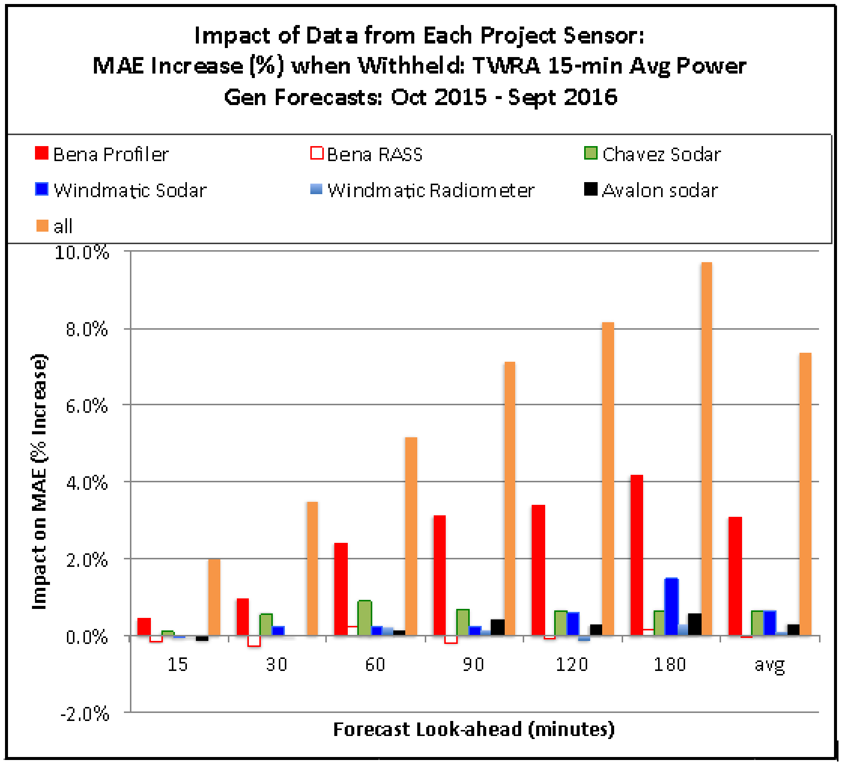
MAE (% of capacity) versus look-ahead time for 0-3 hour forecasts of the 15-minute average wind power production from the TWRA aggregate over the one-year period from October 2015 to September 2016 for each of the five source-dependent sets of predictors (listed in Appendix D) employed in the predictor source category experiment

5.4.4 Contributions of Project Sensors

In addition to the assessment of the aggregate effect of the data from the targeted sensor network, an evaluation of the contributions of each sensor was also performed. This evaluation was done by building a set of XGBoost models that each excluded the use of predictors calculated from a project sensor from the full set of predictors.

The results from these predictor-withholding experiments are depicted in Figure 15. The increase in MAE (relative to forecasts with the full set of predictors) is shown for each sensor. The largest effect was from the Bena wind profiler, which by itself achieved 71% of the overall MAE reduction. This sensor was the farthest upstream from the TWRA and also provided data over the deepest layer. The Chavez and Windmatic sodars had smaller effects of about 0.7%. The effect of the Avalon sodar was near 0.3%, while the other sensors had little to no effect. The total benefit from adding all project sensors (7.34%) was greater than the sum of the benefits from adding any sensor to the remaining sensors (4.82%).

Figure 15: Impact of Data From Each Project Sensor: MAE Increase When Withheld



Increase in MAE by look-ahead time for 0-3 hr forecasts of the 15-minute average TWRA aggregate wind power generation over the one-year period from October 2015 to September 2016 when all predictors derived from the data of each sensor (non-orange columns) were removed from the predictor pool and when predictors based on any project sensor data were removed from the set of predictors (orange columns).

5.4.5 Operational System Configuration: Design and Performance

In a true operational mode, the user expects a forecast to be delivered for every forecast cycle. Therefore, a contingency plan had to be developed to enable the delivery of the best possible forecast even when only subsets of the full dataset are available. The research team developed the forecast contingency plan by creating a hierarchy of backup forecasts to provide the best possible forecast given the data available at any one time while limiting the number of separate backup forecast configurations to six for manageability.

To provide the best forecast possible, five backup configurations were created as shown in Table 7. The first few excluded only the data with the greatest availability issues to provide as much skill as possible. The final backup (the forecast of last resort) used only time data and had 100% availability. However, the data necessary to be used only rarely (about 1% of the forecast intervals).

Table 7: Very Short-Term Forecast System Specifications for the Primary and Five Backup Configurations

Forecast	Predictors Used	Portion of Total Forecasts	MAE Increase Over Primary Forecast
Primary	Use data from all sensors.	51.87%	0%
Backup #1	Exclude Avalon and Windmatic sodars and Windmatic radiometer.	19.88%	1.65%
Backup #2	Exclude all sensors excluded in Backup #1 plus Bena wind profiler.	18.67%	6.81%
Backup #3	Exclude all sensor excluded in Backup #2 plus pressure difference.	8.84%	9.21%
Backup #4	Include time and onsite power/wind data only.	0.42%	12.15%
Backup #5	Include time only.	1.03%	32.45%

5.5 Forecast Performance Analysis: Ramp Event Prediction

Evaluating the effect of the machine learning tools and the project sensor data on ramp event prediction was based on forecasts of the maximum and minimum 60-minute ramp rate over the 180-minute period following the forecast issue time. All the ramp rates were calculated from the 15-minute average generation. Upward ramps were addressed by generating forecasts for the maximum 60-minute ramp rate occurring entirely within the first 180 minutes of the forecast period (RAMPMAX). RAMPMAX is typically greater than zero but can be negative during periods when the power steadily decreases for three or more hours. Similarly, downward ramps were addressed by producing forecasts for the minimum 60-minute ramp rate occurring entirely within the first 180 minutes of the forecast period (RAMPMIN). RAMPMIN is typically less than zero but can be positive during periods when the power steadily increases for three or more hours.

The forecasts were evaluated from two perspectives: (1) the typical errors in predicting the maximum and minimum ramp rates over all forecast periods during the evaluation year and (2) the ability to identify the occurrence of large ramps defined as cases that had maximum and minimum ramp rates that exceeded specified thresholds. The first perspective is dominated by cases with average ramps that are typically small and caused by minor fluctuations in the winds often associated with small-scale weather features near the TWRA. The second perspective is dominated by the few cases in which large ramp rates are observed. These are the cases for which the value of accurate forecasts is highest for system operators.

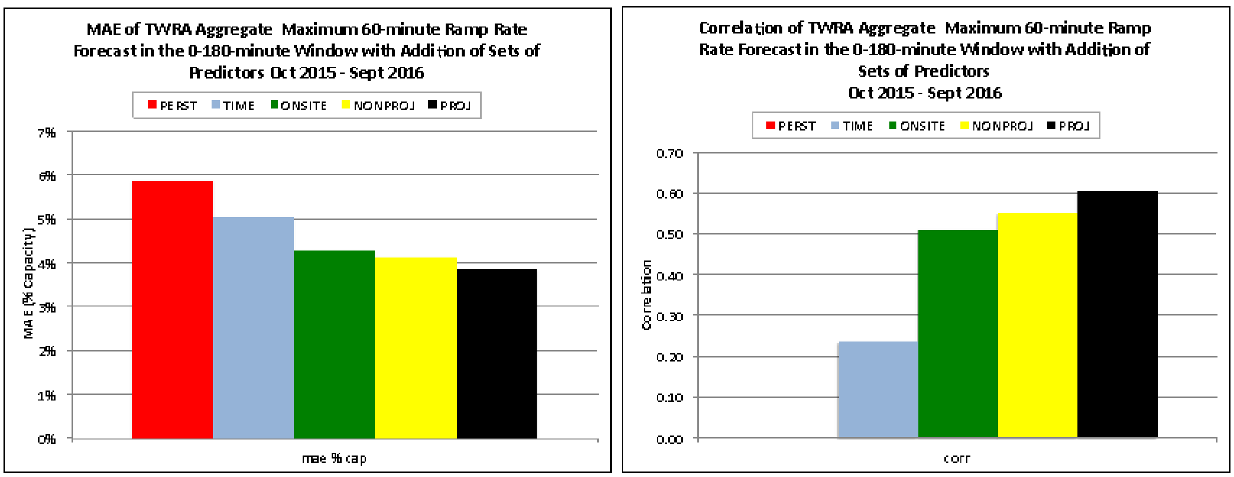
The performance of the maximum and minimum ramp rate forecasts over all forecast cases were evaluated by calculating the MAE and RMSE of the ramp rate forecasts and the correlation and R^2 values between the forecasted and observed (outcome) ramp rates. The ability to predict the occurrence of large ramp events was evaluated by using the critical success index (CSI).

5.5.1 Ramp Rate Prediction

The best and final configuration for the 60-minute ramp rate prediction mode (maximum and minimum ramp rate in the three hours following forecast issue time) produced an MAE of 3.84% of capacity for the maximum ramp rate forecasts (Figure 16) and 3.15% for the minimum ramp rate forecasts (Figure 17). This was an MAE reduction of 34.6% for the maximum ramp rate and 48.0% for the minimum ramp rate relative to a forecast of a zero ramp rate (i.e., persistence). The forecasts explained 36.6% of the variance in the maximum ramp rate and 50.7% of the variance of the minimum ramp rate.

The predictors that were used in the best and final configuration are listed in Appendix D.

Figure 16: Maximum 60-Minute Ramp Rate (RAMPMAX) Forecast MAE and Correlation

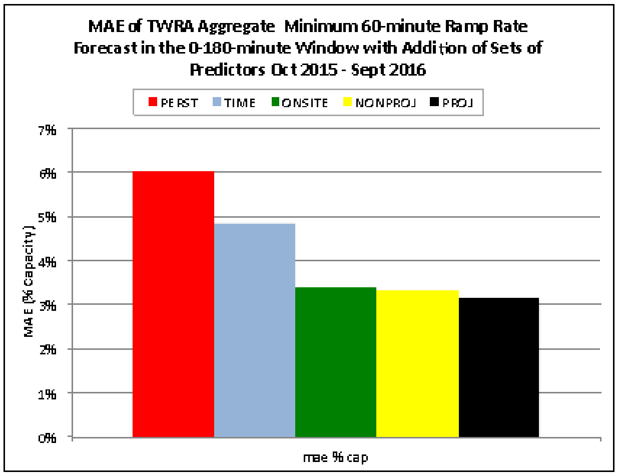


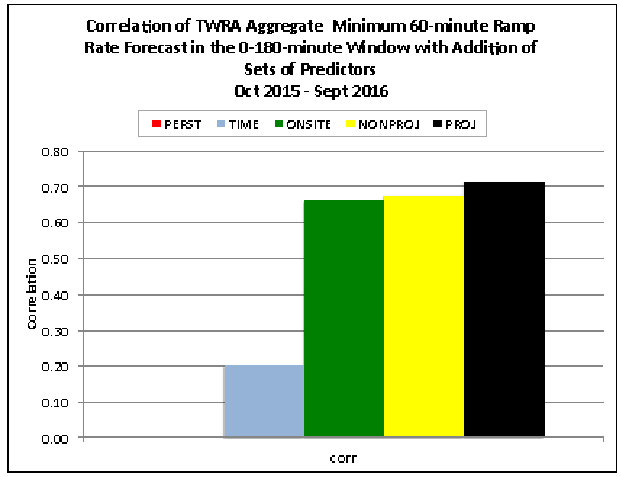
The MAE (left) and forecast vs. observed correlation (right) for 0-3 hour forecasts of the maximum 60-minute ramp rate (RAMPMAX) with an XGBoost model trained with five subsets of predictors. The composition of the predictor subset is cumulative from left to right (i.e., from red to black) with each successive subset to the right including all the predictors of the previous subset plus a set of additional predictors.

5.5.1.1 Contributions of Project Sensors

In addition to assessing the effect of the aggregated project sensor data on the ramp event forecasts, the relative contributions of the data from each sensor and combinations of sensors were also evaluated. The addition of predictors calculated from the data from all project sensors to the pool of all other predictors for the forecasts of the maximum ramp rate produced a 6.9% reduction in the MAE and a 12.4% improvement in the correlation coefficient. For the minimum ramp rate, the addition of the predictors from the project sensor data yielded a 5.3% reduction in MAE and an 11.4% improvement in the correlation coefficient. These results indicate that the project sensor data had slightly more benefit for the predicting upward ramp rates than downward ramp rates.

Figure 17: Minimum 60-Minute Ramp Rate (RAMPMIN) Forecast MAE and Correlation With Observation





The MAE (left) and forecast vs. observed correlation (right) for 0-3 hour forecasts of the minimum 60-minute ramp rate (RAMPMIN) with an XGBoost model trained with five different subsets of predictors. The composition of the predictor subset is cumulative from left to right (i.e., from red to black) with each successive subset to the right including all the predictors of the previous subset plus a set of additional predictors.

Table 8 lists the project sensors in order of associated benefit for the predicting the maximum upward and downward ramp rates. The most valuable sensor for both the prediction of upward and downward ramp rates was the Bena radar wind profiler. The Windmatic sodar provided substantially more benefit for upward ramps than for downward ramps while the Chavez sodar and Windmatic radiometer provided substantially more benefit for downward ramps than for upward ramps.

Table 8: Sensors Ranked in Order of Contributions

Rank	Upward Ramps (sensor, % MAE benefit added, % of correlation benefit added)	Downward Ramps (sensor, % MAE benefit added, % of correlation benefit added)
1	Bena profiler (78.3%, 88.9%)	Bena profiler (61.3%, 64.9%)
2	Windmatic Sodar (16.5%, 5.5%)	Chavez sodar (17.7%, 21.6%)
3	Chavez Sodar (3.9%, 5.6%)	Windmatic radiometer (9.0%. 10.8%)
4	Avalon Sodar (0.8%, 1.8%)	Windmatic sodar (9,2%, 8.1%)
5	Windmatic Radiometer (2.7%, -3.7%)	Avalon Sodar (4.2%, 0.0%)

Based on MAE and correlation coefficient metrics) to improving the prediction of upward and downward ramp rates.

5.5.2 Prediction of Large Ramp Events

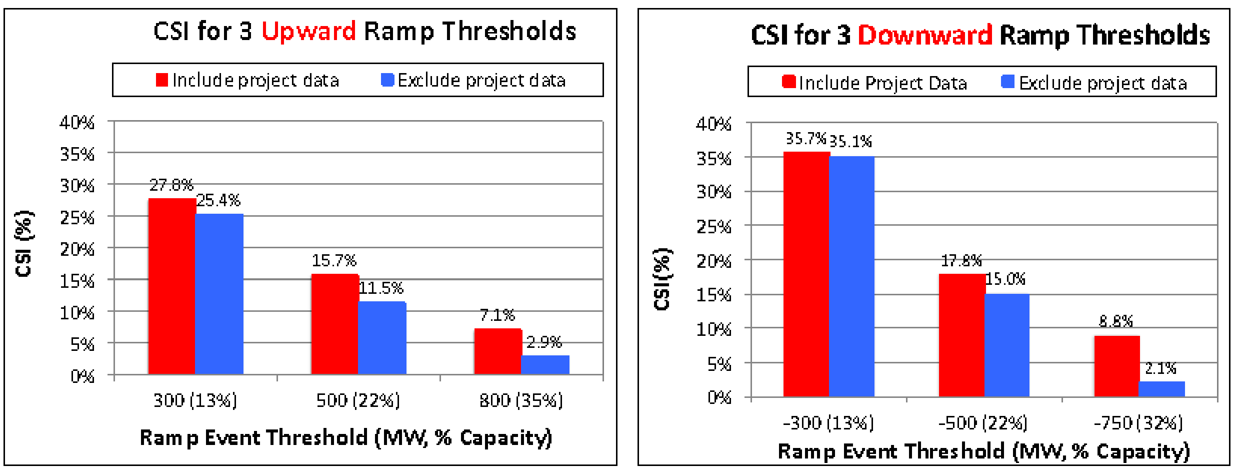
One of the most important operational aspects of the ramp rate forecasts for grid management applications is the ability of these forecasts to accurately anticipate large ramp events by providing warning for as many events as possible while minimizing the number of false alarms.

The ability of the forecast system to address this objective was evaluated through the use of the previously defined CSI metric.

The CSI was computed by selecting three thresholds for RAMPMAX (300, 500 and 800 MW in 60 minutes) and RAMPMIN (-300, -500 and -750 MW in 60 minutes). The largest threshold was reduced slightly for RAMPMIN to ensure an adequate sample size.

The CSI values for the large upward and downward ramp event forecasts are shown in Figure 18. The overall performance is better for smaller ramps (i.e., the 300 MW event threshold) for both upward and downward ramps. A comparison of the blue and red columns in the two charts indicates the impact of the project sensor data on the forecasts of the large ramp events. The CSI score for the forecasts with the project sensor data are slightly higher for the 300 MW threshold for upward and downward ramps. However, the CSI scores are much higher for the forecasts with the project sensor data for the highest thresholds. **These results indicate the project data provide more benefit for the prediction of larger ramps.**

Figure 18: CSI for Ramp Event Forecasts With and Without Project Sensor Data



Critical success index (CSI) for forecasts of upward (left) and downward (right) 60-minute ramps for the 2319 MW aggregate of TWRA wind generation, which are defined by three event thresholds for upward (300 MW, 500 MW and 800 MW) and downward (-300 MW, -500 MW and -750 MW) ramps. The blue columns depict the CSI for forecasts without the use of project sensor data and the red columns show the CSI for forecasts that used the project sensor data. Higher scores indicate better performance.

5.6 Conclusions

Overall, the machine-learning-based very short-term prediction model exhibited considerable skill in the 0-3 hour prediction of the time series of the 15-minute average power production, the maximum and minimum ramp rates and the occurrence/nonoccurrence of large ramps. The project sensor data contributed to substantial improvements in the performance of all three forecast modes.

CHAPTER 6: Wind Ramp Forecast System Evaluation

6.1 Introduction

Task 6 is the capstone activity of the project. It was designed to integrate all of the individual improvements developed and implemented during the project and determine the composite effect of the improvements relative to a baseline forecast system. The experiment to assess the improvements was given the name "Forecast Improvement Assessment Experiment" (FIAE).

A number of forecast improvements were developed and implemented within the tasks of this project. These improvements were:

- Using a targeted sensor network (Task 3).
- Customizing a Numerical Weather Prediction (NWP) model configuration for 0-15 hour ahead wind forecasting in Tehachapi Pass (Task 2).
- Adapting and refining a GSI-EnKF hybrid data assimilation method to assimilate targeted sensor network data into a high-resolution NWP model (Task 4).
- Developing and applying a machine-learning-based statistical time series forecast model to produce 15-minute updates of 0-3 hour forecasts (Task 5).
- Developing and applying a machine-learning-based Model Output Statistics (MOS) procedure for each NWP model (initial component of Task 6).

All these forecast system improvements were integrated into a multi-method ensemble forecast system. The forecasts produced by this type of system are a composite of individual forecasts from a set of different forecast methods.

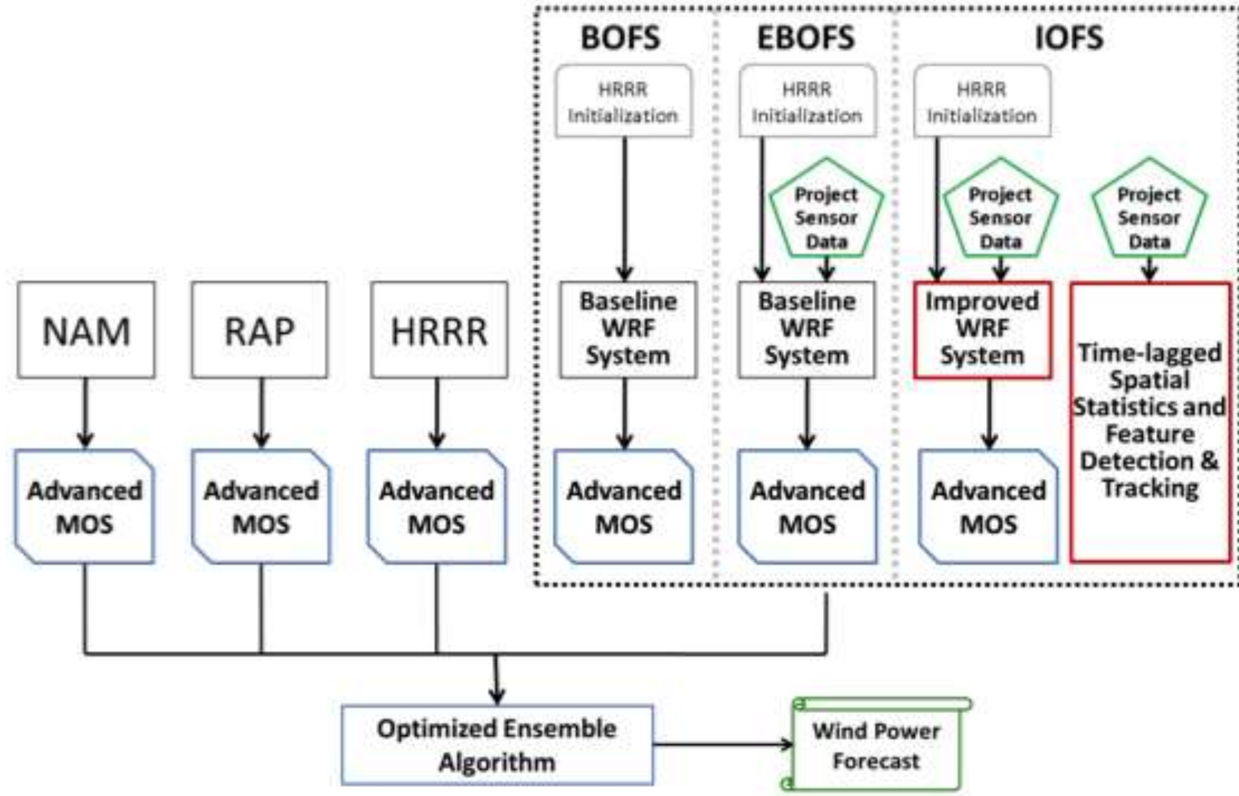
6.2 Experimental Design

6.2.1 Basic Structure

The basic structure of the FIAE is a comparison among four different versions of a multi-method ensemble forecast system. The underlying concept is that current state-of-the-art forecasts are produced by combining forecasts from a set of different physics-based and statistical methods. Therefore, the most appropriate way to assess the value of a new or improved method is assessing the effect of adding the improvements to a system based upon a set of existing state-of-the-art methods. In this case, the baseline system consisted of a set of three National Weather Service (NWS) NWP models whose output was statistically processed by an advanced machine-learning algorithm. Three other versions of the system were created by adding project-based systems to this core set of NWS-based methods.

A schematic of the interrelationships among the four ensemble-based forecast systems is shown in Figure 19. The details of each component employed in these systems are documented in Appendix F.

Figure 19: Structure Schematic of the FIAE



A schematic depiction of the components and data flow of the Forecast Improvement Assessment Experiment (FIAE),

The first system, labeled "NWS," is the previously noted baseline system that uses only the three NWS NWP models and the machine-learning-based processing of the output. This system is depicted by the components that are outside the dashed box. The three NWP models in this system are (1) the North American Mesoscale (NAM) Model, (2) the Rapid Refresh (RAP) model and (3) the High-Resolution Rapid Refresh (HRRR) model. An advanced Model Output Statistics (MOS) procedure is applied to the output of each of these models. The MOS procedure statistically transforms the NWP forecast variables directly to predictions of the power production of individual wind generation facilities in the TWRA. The power production forecasts for the facilities are then combined to produce the aggregate of all TWRA facilities considered in this project and six subaggregates of those facilities. The power generation forecasts from each of these three NWP-MOS systems are then combined via the Optimized Ensemble Algorithm (OEA) to produce the final forecast from this ensemble. The OEA is also based on an advanced machine learning algorithm that constructs an optimal composite of the three input forecasts.

The second ensemble system consisted of the NWS-based component in the first system plus the simplest version of the project subsystems is labeled "BOFS" (baseline operational forecast system). This ensemble was constructed to assess the effect of the customized version of the Weather Research and Forecasting (WRF) model (Skamarock et al, 2008) that resulted from the sensitivity experiments conducted in Task 2. In the FIAE, the BOFS-WRF system was initialized with data from the RAP model and it also used lateral boundary condition data from the RAP. There was no assimilation of data into the initial state extracted from the RAP dataset. That is, the data from the sensor network of the project was not used and neither was any other local data (such as the meteorological data from the wind generation facilities) that is not used by the RAP model initialization.

The third ensemble system was the same as the second system except the data from the targeted sensor network of the project were assimilated into the initial state used for the WRF forecasts using the GSI data assimilation method. This system called the *enhanced baseline operational forecast system* (EBOFS). This system was designed to assess the effect of assimilating data from the targeted sensor network of the project.

The fourth ensemble system was designed to assess the integrated effect of the forecast system improvements implemented and refined in this project. It included the three NWS NWP model components as in the other three ensemble systems. The NWP portion of this ensemble also included the custom configured version (from Task 2) of the WRF model that was used in the BOFS and EBOFS ensembles. The data assimilation component of the WRF system, however, was changed to the hybrid data assimilation system configured and refined by the UC Davis group in Task 4 of this project. This method was used to assimilate the data from the targeted sensor network. In addition to the different data assimilation method, the IOFS ensemble also included the non-NWP very short-term statistical forecast method developed in Task 5 of this project. This method produced forecasts only for the 0-3 hour look-ahead period.

The forecasts from the NWS, BOFS and EBOFS ensembles were evaluated over a one-year period that extended from October 1, 2015 to September 30, 2016. Forecasts from each ensemble were generated on a six-hour cycle during this period. Ideally, IOFS forecasts would also have been produced for the same cycles during this one-year period. However, since the production of the IOFS forecasts had to wait until all method refinement work in the project was completed, there were not sufficient time and resources available to generate and evaluate a full year of IOFS forecasts. The sample of forecasts from the IOFS ensemble, therefore, was limited to six months. The six months in this sample were February, March, May, July and September of 2016. These months were selected via the application of two criteria: (1) representation of each of the three primary Tehachapi seasonal wind regimes (mid-latitude storms, diurnal and monsoon) and (2) above-average number of large wind ramp events and generally higher-than-average wind variability. The corresponding months from the NWS, BOFS and EBOFS forecasts were used for performance comparisons with the IOFS forecasts.

6.2.2 Forecast Assessment Plan

A difficult issue in this type of project is how to assess the effect of the forecast system improvements on forecast performance in a meaningful way. There are two core issues that give rise several secondary issues. One core issue is the selection of metrics to assess forecast performance. The choice of metrics determines the attributes of the forecast system that is evaluated. Alternate choices of metrics can often lead to very different perspectives on the performance of the same forecast system. Ideally, the selected metrics should measure the way in which the target application is sensitive to forecast error. This is often difficult to quantify and/or the application is a composite of several use cases with different patterns of sensitivity to forecast error. The second core issue is the selection of an appropriate baseline forecast to serve as the basis for assessing forecasting improvement (i.e., improved with respect to what?). Ideally, the demonstrated improvement is with respect to a state-of-the-art forecast. At any point in time, however, it is often difficult to determine "state-of-the-art" performance.

Because of project time and resource constraints, the focus was on the forecasts performance of the time series for the 15-minute average TWRA aggregate power production. Since the set of target applications for this project was broad, it was decided to employ the widely used and standard metrics of mean error (ME), or bias, the mean absolute error (MAE) and the root mean square error (RMSE). Certainly, the forecasts produced in this project could be evaluated with other metrics and that might lead to different conclusions. However, using statistical models in the forecast system implies the specification of a performance objective and that this specification of the objective should not be decoupled from the demands of the application or the evaluation protocol. That is, the system should be optimized for needs of the application and how the performance will be evaluated.

6.3 Forecast Performance Analysis

The forecasts of the time series of the power generation were analyzed from two perspectives: (1) the performance of the baseline system and related components and (2) the effect of adding the improvements implemented in the IOFS.

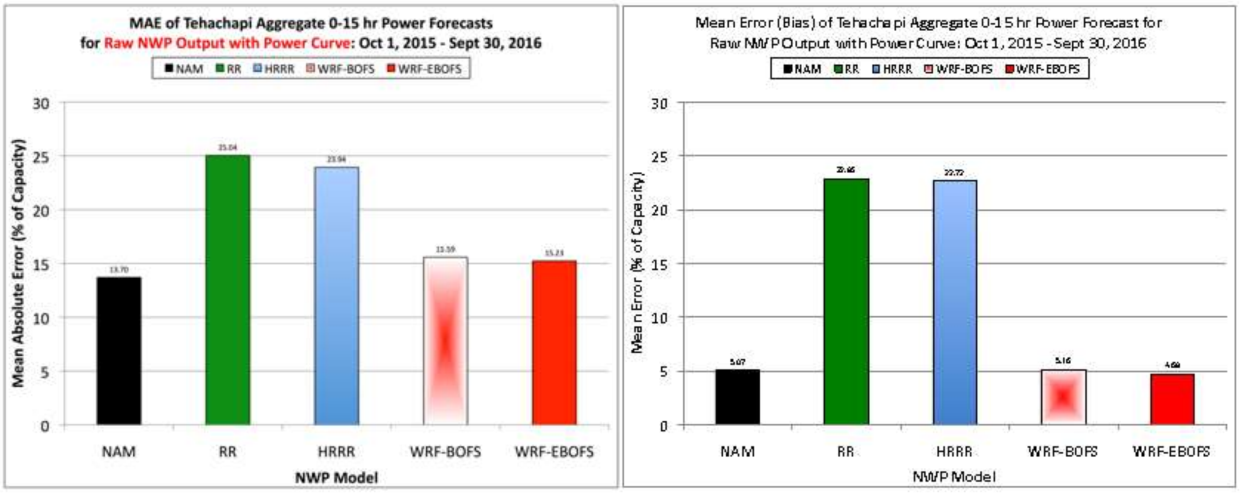
6.3.1 Baseline Performance

The performance of the baseline system was analyzed on three levels: (1) the raw forecasts of the NWP models, (2) the MOS-adjusted NWP forecasts and (3) ensemble composite forecasts.

6.3.1.1 Raw NWP

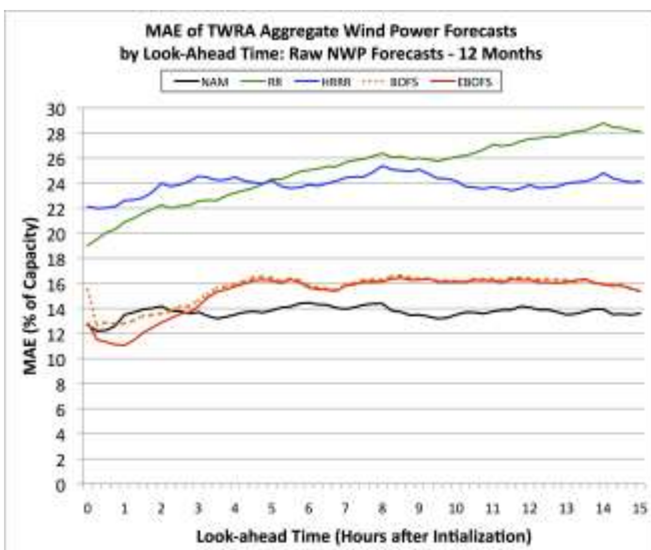
The MAE and bias for each of the baseline NWP systems over the one-year evaluation period are shown in Figure 20. The best forecast among the baseline models was produced by the National Weather Service's NAM model. It achieved an MAE of 13.7 % of capacity for the predictions of the 15-minute average power production over all the forecast intervals (60) in the full 15-hour forecast period. The second lowest MAE over the full 15-hour period of 15.2% of capacity was produced by the WRF-EBOFS model. This was the project NWP system that included the assimilation of the project sensor data. The WRF-BOFS system, which was identical to the WRF-EBOFS system except that it did not assimilate the project sensor data, achieved an MAE of 15.6% over the entire 15-hour period. Thus, the assimilation of the project sensor data with the basic data assimilation scheme used in EBOFS produced a 2.3 % reduction in MAE over the 15-hour period. However, an examination of the MAE by look-ahead time that is depicted in Figure 21 indicates that almost all the MAE reduction of WRF-EBOFS vs. WRF-BOFS is achieved in the first three hours of the forecast period. The maximum percentage MAE reduction by WRF-EBOFS of 13.6 % (i.e., from 12.8 % of capacity to 11.1 % of capacity) is achieved at a look-ahead time of one hour. After the three-hour look-ahead time, the MAE of the two systems is almost the same.

Figure 20: MAE and Bias of Baseline NWP Power Generation Forecasts



Mean absolute error (MAE; left) and mean error (bias; right) of the 0-15 hour raw NWP forecasts of the time series of the 15-minute average of the TWRA aggregate power production for four forecast cycles per day for the one-year period extending from October 1, 2015 to September 30, 2016.

Figure 21: MAE of Baseline NWP Power Generation Forecasts by Look-Ahead Time



Mean absolute error (MAE) by look-ahead time of the 0-15 hour raw NWP forecasts of the time series of the 15-minute average of the TWRA aggregate power production for 4 forecast cycles per day for the one-year period extending from October 1, 2015 to September 30, 2016.

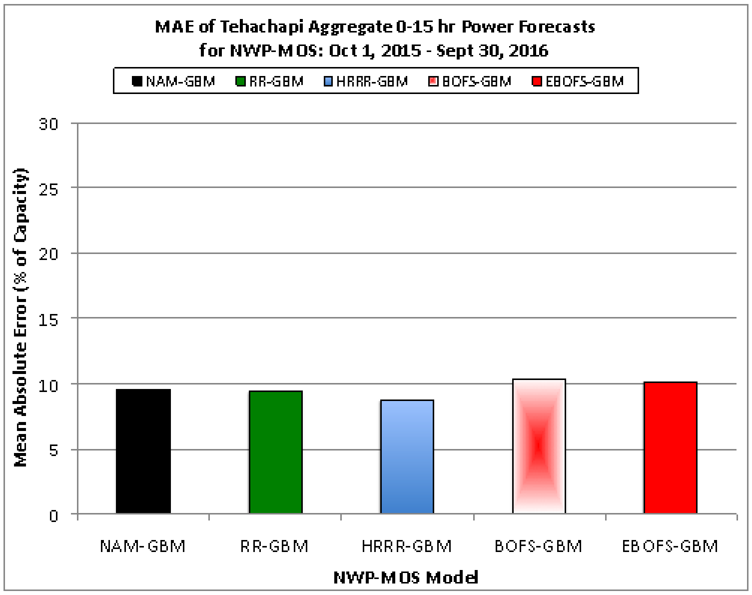
The worst performance among the raw NWP forecasts was produced by the two rapid update NWP models: the RAP and HRRR. The overall MAEs of 25.0% and 23.9%, respectively, were much worse than the three other NWP systems employed in the project. However, an examination of the full period bias of the raw NWP forecasts indicates that both of these systems produced forecasts with a very large positive bias. This large positive wind speed forecast bias was noted for the RAP forecasts in the analysis of the NWP sensitivity experiments in Task 2. Although it was not analyzed in Task 2, the HRRR model as a higher spatial resolution version of the RAP system also had a similar bias pattern. The other three models also produced power generation forecasts that had a positive bias, but the magnitude of the bias was substantially lower than the size of the bias in the RAP and HRRR forecasts.

The significance of the forecast bias is that systematic errors such as those indicated by the bias statistics can often be substantially reduced by statistical postprocessing such as MOS. Thus, high MAE values are not necessarily an indication that a forecast method will not have much value within an ensemble of forecasts. However, it does indicate that the forecast cannot be directly used without statistical postprocessing to significantly reduce the bias.

6.3.1.2 MOS-Adjusted NWP

The MAE of the MOS-adjusted 0-15 hour NWP forecasts from each modeling system over the full one-year FIAE evaluation period is shown in Figure 22. The HRRR-MOS system achieved the lowest MAE of 8.75% of capacity. As noted in the previous section, the MAE of the raw HRRR forecasts was close to the worst among the five modeling systems, with only the raw forecasts from the RAP having a slightly higher MAE. However, the MOS procedure reduced the MAE of the HRRR forecasts by about 63.5%. This was the largest reduction of all the NWP models. This large reduction in MAE was associated with the reduction in the very large positive bias that was present in the raw HRRR forecasts. The three modeling systems (NAM, WRF-BOFS and WRF-EBOFS) which had produced raw forecasts with a much smaller positive bias received less benefit from the MOS procedure. The amount of benefit provided by the MOS procedure was not fully explained by the magnitude of the bias of the raw NWP forecasts. While the magnitude of the bias is a major factor in determining the benefit obtained from the MOS procedure, there are other factors that play a role.

Figure 22: MAE of Baseline NWP-MOS Power Generation Forecasts



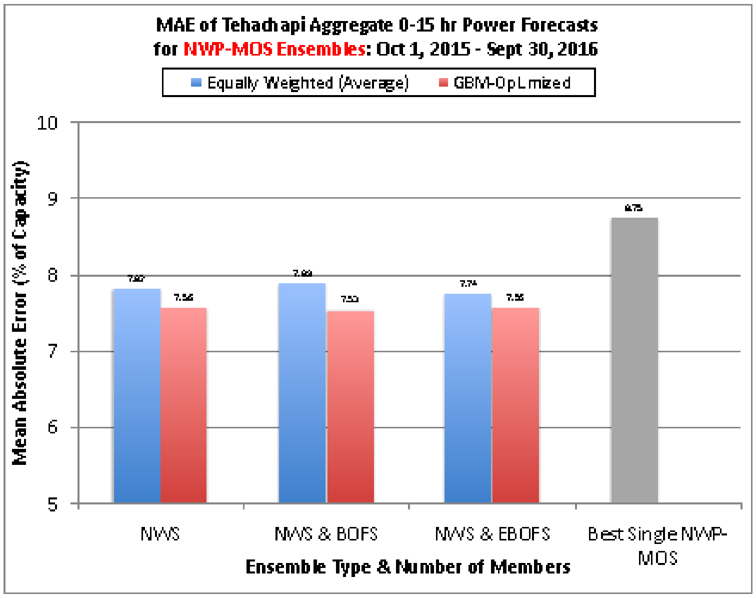
Mean absolute error (MAE) of the 0-15 hour NWP-MOS forecasts of the time series of the 15-minute average of the TWRA aggregate power production for four forecast cycles per day for the one-year period extending from October 1, 2015 to September 30, 2016.

The bottom line from evaluating the NWP-MOS forecasts was that the application of the machine-learning-based MOS procedure resulted in a large reduction in the MAE of the 0-15 hour power production forecasts and that the reduction was strongly (but not exclusively) linked to the magnitude of the bias in the raw NWP forecasts from each model. After the application of the MOS algorithm, the best performing NWP system was the HRRR-MOS system with an overall MAE of 8.75 % of capacity.

6.3.1.3 NWP-MOS Ensembles

The MAE of forecasts from each of the three ensembles and, for a reference point, the best performing NWP-MOS system is shown in Figure 23. The blue columns depict the MAE for an equally weighted composite while the red columns show the MAE for a machine-learning optimized composite. The MAEs for the three composites are very similar and in each case the optimized composite performed better than the equally weighted counterpart. The MAE reduction associated with the optimized composite ranged from 2.4% to 4.5%. However, all the composites performed substantially better than the best model system. The average MAE reduction relative to the best system was about 10% for the equally weighted composites and 13% for the optimized composites. The bottom line was that neither of the initial project-based composites outperformed the NWS-only composite.

Figure 23: MAE of Baseline NWP-MOS Ensemble Power Generation Forecasts



Mean absolute error (MAE) of the 0-15 hour raw NWP-MOS ensemble forecasts of the time series of the 15-minute average of the TWRA aggregate power production for 4 forecast cycles per day for the one-year FIAE period.

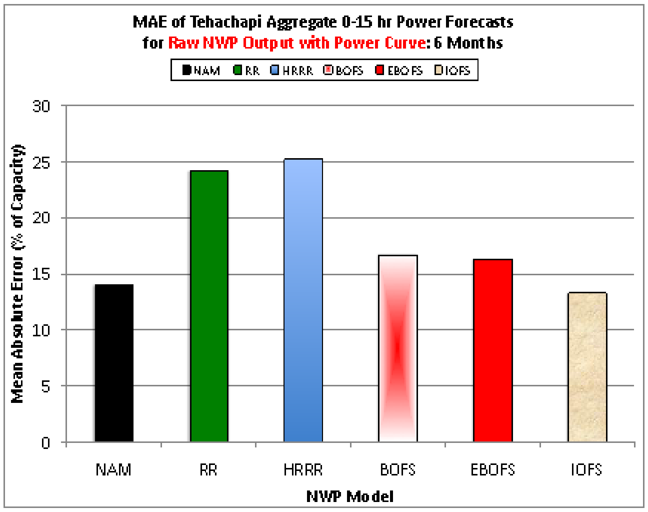
6.3.2 Impact of the IOFS

This forecast improvement assessment task (Task 6) determined the integrated efect of the forecast system improvements implemented and refined in this project. This was done by generating a set of forecasts for a portion of the one-year assessment period with the IOFS version of the forecast system, which included all of the improvements implemented in this project.

6.3.2.1 Raw NWP

A comparison of the MAE of the power production forecasts derived from all of the NWP systems over the six-month is shown in Figure 24. This chart is analogous to the one depicted in the left panel of Figure 20, except that it is for a six-month subsample and the MAE of the WRF-IOFS system has been added. Since the chart is for a six-month period, the MAE values for the NWP models that are also shown in Figure 20 are not the same in Figure 24. However, they are not dramatically different and the relative performance is quite similar. This indicates that the performance patterns for the six-month subsample are fairly similar to those for the full 12-month sample. Thus, it is fairly likely that statistically significant conclusions drawn from the six-month sample also apply to the 12-month sample.

Figure 24: MAE of IOFS and Baseline NWP Power Generation Forecasts



Mean absolute error (MAE) of the 0-15 hour WRF-IOFS and other raw NWP baseline forecasts of the time series of the 15-minute average of the TWRA aggregate power production for four forecast cycles per day for the six-month FIAE subsample.

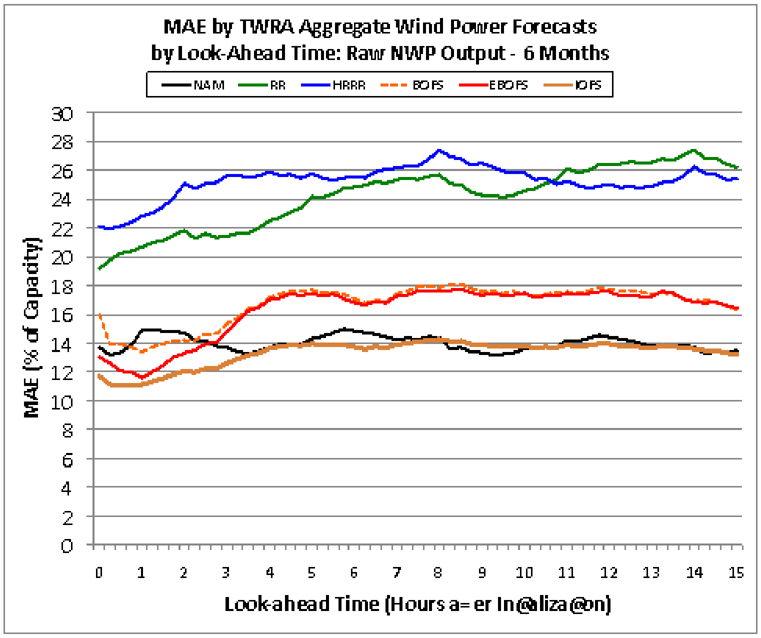
The WRF-IOFS system produced the lowest MAE for the full 15-hour forecast period of all of NWP systems over the six-month period. It yielded an MAE of 13.3% of capacity. This was 4.8% lower than the 14.0% of capacity MAE produced by the National Weather Service's NAM model for the six-month period. The NAM model was the best performing system among the five baseline models (i.e., without the WRF-IOFS) over the full 12-month period and it also produced the second lowest MAE (behind the WRF-IOFS) among the six models in the six-month subsample.

The temporal pattern of the MAE for each of the six raw NWP forecasts over the 15-hour forecast period is shown in Figure 25. This chart indicates that the performance advantage of the WRF-IOFS system over the NAM system was mostly in the first three hours of the forecast period. The largest percentage reduction in the MAE by the WRF-IOFS relative to the NAM was for the one-hour look-ahead time. After that time, the MAE of the raw forecasts from the NAM and WRF-IOFS are very similar.

In contrast, the reduction in MAE relative to WRF-BOFS and WRF-EBOFS extended throughout the 15-hour forecast period. The improvement over WRF-BOFS was fairly uniform whereas the improvement over the WRF-EBOFS system was less during the first three hours of the forecast period and greater during the latter portion of the period. The raw WRF-IOFS forecasts achieved a lower MAE even though they had a somewhat higher positive bias than the WRF-BOFS, WRF-EBOFS and NAM forecasts. The presence of a higher bias means that there is more opportunity for error reduction in the statistical postprocessing component of the forecast system.

The team concluded the results from the comparison of the raw forecasts from the WRF-IOFS performance relative to the raw forecasts from the other NWP system was that the WRF-IOFS system substantially improved upon the performance of all the other five systems and was the best performing system for the six-month subsample.

Figure 25: MAE of IOFS and Baseline NWP Power Generation Forecasts by Look-Ahead Time



Mean absolute error (MAE) by look-ahead time of the 0-15 hour raw baseline and IOFS NWP forecasts of the time series of the 15-minute average of the TWRA aggregate power production for 4 forecast cycles per day for the six-month FIAE subsample.

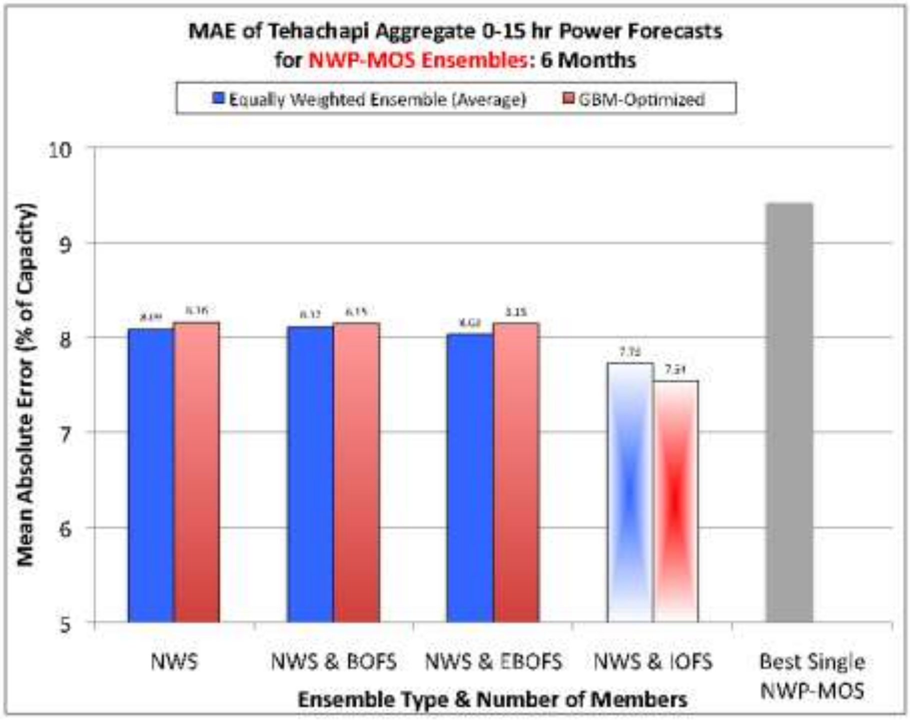
6.3.2.2 NWP-MOS Ensembles

While the team was encouraged that the forecasts produced directly from the WRF-IOFS NWP system improved upon the performance of all of the other NWP systems in the project, the ultimate test of the value of the improvements implemented in the IOFS NWP system is the effect on the forecast performance when the forecasts produced by this system are processed by a MOS procedure and included in an ensemble composite. This is how these forecasts would be used in an operational forecast system unless the IOFS NWP-MOS forecasts were so vastly superior to the other NWP forecasts that they outperformed the best ensemble composite.

The MAEs for the IOFS NWP ensemble and the three reference NWP ensembles as well as the best single NWP-MOS system for the six-month subsample are shown in Figure 26. It should be noted that the MAEs for the reference systems shown in Figure 26 are different from those depicted in Figure 23 since these are for a six-month period and the data in Figure 23 are for the full 12-month sample. The MAEs for the ensembles of the reference systems are somewhat higher (about 3.5% higher for the equally weighted ensembles) for the six-month period than for the 12-month period. In the case of the equally weighted ensemble composites this is due to the higher forecast difficulty level for the six months in the subsample. The months in the six-month sample were selected to have a higher-than-average number of large ramps and this was associated with a higher-than-average amount of wind variability in general.

Another feature of note in the six-month MAE pattern of the reference systems is that the optimized ensembles actually had higher MAEs than their equally weighted counterparts. This is the opposite of the pattern seen in the 12-month results (Figure 23) in which the optimized ensemble produced somewhat lower MAE values. This may be attributable to the shorter data sample (5 months vs. 11 months) used to train the statistical models with the 6-month sample.

Figure 26: MAE of IOFS and Baseline NWP-MOS Ensemble Forecasts

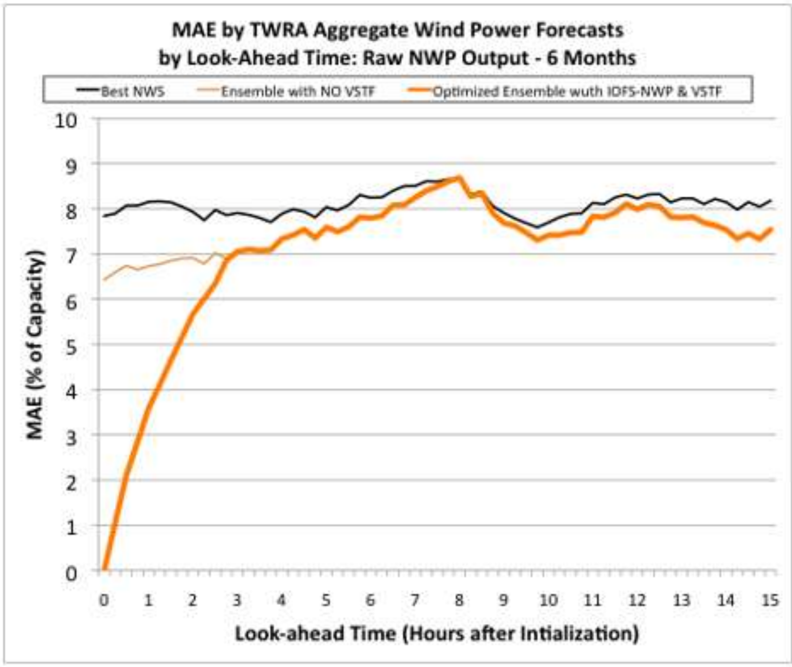


Mean absolute error (MAE) of the 0-15 hour NWP-MOS ensemble composite forecasts of the time series of the 15-minute average of the TWRA aggregate power production for four forecast cycles per day for the six-month FIAE subsample.

The IOFS ensemble produced a substantial MAE reduction relative to the three reference ensembles. The equally weighted IOFS ensemble yielded nearly a 4.5% lower MAE over the 0-15 hour forecast period than the equally weighted composites from the NWS ensemble, which was the fundamental baseline for this project. Interestingly, although the ensemble optimization approach did not lower the six-month MAE for any of the reference ensembles, it did lower the MAE for the IOFS ensemble. This resulted in a 6.7% IOFS-induced improvement in the MAE over the entire 15-hour forecast period relative to the MAE of the best NWS ensemble (i.e., the equally weighted one).

The dependence of this overall (i.e., 0-15 hour) 6.7% MAE reduction on the forecast look-ahead time is shown in Figure 27. This chart depicts the MAE of the best composite forecast from the NWS NWP-MOS ensemble and the optimized IOFS NWP-MOS ensemble (which of course includes the components of the NWS ensemble). The data depicted in this chart indicate that a large fraction of the 6.7% improvement occurs in the first four hours of the forecast period. This is most likely because much of the improvement is associated with the assimilation of the project sensor data and that has the greatest effect in the early part of the forecast period. However, there is some improvement over almost all of the 15-hour period.

Figure 27: MAE of IOFS and Baseline NWP-MOS Forecasts by Look-Ahead Time



Mean absolute error (MAE) of the 0-15 hour raw NWP-MOS Ensemble forecasts of the time series of the 15-minute average of the TWRA aggregate power production for four forecast cycles per day for the six-month FIAE subsample.

6.3.2.3 NWP-MOS and VSTF Ensemble

The MAE of an IOFS ensemble composite forecast that included the VSTF forecasts is shown in Figure 27 along with the baseline NWS NWP-MOS ensemble and the IOFS NWP-MOS ensemble. The addition of the VSTF forecasts to the ensemble dramatically reduced the MAE in the first two hours of the forecast period. A large portion of this reduction is associated with the information contained in the recent history of power generation and meteorological conditions at the wind generation facilities. However, a substantial amount is attributable to the predictive information contained in the data from the project sensors. The effect of the VSTF on the performance of the IOFS ensemble essentially disappears about 15 minutes before the end of the VSTF three-hour forecast period. This performance pattern is expected for a short-term non-NWP statistical forecast model and is the fundamental reason why this method was designed to produce forecasts for only the first three hours of the 15-hour forecast period.

The MAE over the full 15-hour forecast period due to the addition of forecasts from the VSTF model decreased from 7.54 % of capacity with the non-VSTF version of the IOFS ensemble to 7.02% of capacity. The net result was a 13.5% reduction in MAE relative to the baseline NWS NWP-MOS ensemble. As seen in Figure 26, a substantial portion of the MAE reduction from the NWP data assimilation improvements overlapped with the MAE reduction from the use of the VSTF model for the first 2.5 hours of the forecast period. However, as noted, there was also some synergy between these two improvements.

6.4 Conclusions and Potential Next Steps

6.4.1 Conclusions

The baseline optimized NWS-MOS ensemble produced an MAE of 7.56% for the forecasts of the 15-minute average power production over the full 15-hour forecast period for the entire one-year evaluation period. Within this ensemble, the best performing NWP-MOS system was the HRRR-MOS, which had an overall MAE of 8.75%. Using an ensemble of NWS models resulted in a 13% MAE reduction relative to the best single model.

The two initial project forecast systems (the BOFS which added a custom-configured NWP model and the EBOFS which added a basic assimilation of data from the targeted sensor network project to the BOFS configuration) that were added to the NWS ensemble did not improve the performance of the optimized NWS ensemble over the full 15-hour forecast period.

Over a six-month period, the IOFS ensemble reduced the 0-15 hour MAE of TWRA aggregate power generation forecasts from the baseline NWS ensemble by 6.7% without the VSTF component and 13.5% with the use of the VSTF. Most of the improvement was in the 0-3 hour portion of the 15-hour forecast window. The concentration of the forecast accuracy improvement in the 0-3 hour period was most likely attributable to the reduction in forecast error was associated with the effective use of data from the project sensor network, which was close (40 km or less) to the wind generation resources and therefore would be expected to provide most of the benefit on a short time scale.

6.4.2 Potential Next Steps

This section presents suggestions for potential next steps to further enhance the value and broaden the applicability of the work done in this project. The most basic opportunity to extend the scope of the effort in this project is the expansion of the evaluation of the impact of the addition of the forecast system improvements in the IOFS to a full year. All of the data is available for this activity. This would make the results of the evaluation experiment more robust since a broader sample in each season could be analyzed.

The second basic opportunity is to evaluate the effect of the IOFS improvements on the performance of ramp event forecasts. The underlying data for this exercise were generated in this project but the data was beyond the scope of the project because of limited project resources.

A third possibility is to conduct NWP experiments to analyze the effect of each project sensor on the NWP forecasts. This was done for the statistics-based VSTF model and those results are presented in the Task 5 report. The analysis of the impact of each sensor on the NWP forecasts would require the execution of a number of NWP forecasts (omitting the assimilation of one sensor in each forecast) for each case under consideration. It would likely be impractical to do this for a one-year or even a six-month period. However, a 30-case sample in each of the three primary Tehachapi Pass weather regimes would be a reasonable objective.

CHAPTER 7: Conclusions and Recommendations

7.1 Wind Power Forecasting

The modeling component of this project began with experiments to evaluate the sensitivity of numerical weather prediction (NWP)-based wind power forecasts for the Tehachapi Wind Resource Area (TWRA) to the configuration of the NWP model. The NWP experiments consisted of 11 runs of the WRF model and a comparison with the forecasts from the National Weather Service's (NWS) operational RAP model for 30 TWRA wind ramp events in late 2014 and 2015. The baseline model configuration was one that has been widely used in California and elsewhere.

Three groups of sensitivity experiments were carried out: (1) six model runs that varied one or more of the submodels relative to the baseline run but did not use the wind turbine drag submodel; (2) three model runs that used the wind turbine drag submodel with different configurations of the other submodels; and (3) two model runs with different resolutions that used the configuration employed by the RAP model.

Analysis of the time series of the average 15-minute hub-height wind speed and power production for the TWRA aggregate using the bias, mean absolute error (MAE) and root mean square error (RMSE) metrics produced three broad conclusions:

- The NWS RAP forecast and all WRF-based forecasts that did not use the turbine drag submodel had a very large positive bias (i.e., the forecast value was too high) for the hub-height wind speed and the power generation.
- The bias and MAE of the power production forecasts exhibited variations of 31% and 19%, respectively, among the forecasts without the turbine drag submodel. These results indicate that the choice of submodels can affect forecasts significantly.
- Using a turbine drag submodel effectively eliminated the positive bias in the wind speed and power production forecasts.

Based on the bias, MAE and RMSE statistics, the best performing model over the 30-case sample was the WRF model that employed the turbine drag submodel as well as several other submodels that were different from those that were employed in the baseline configuration.

Evaluating ramp event forecasts yielded a different perspective on forecast performance. The ramp event forecasts were evaluated with the critical success index (CSI) metric that combines hit, miss and false alarm data for ramp forecasts. Based on the CSI, the forecasts without the turbine drag submodel performed the best. The addition of the turbine drag submodel degraded the ramp event forecast performance and produced a 30% lower CSI score along with a significant (22%) underprediction of the number of ramps. The forecasts from the RAP model configuration produced an even lower CSI score and a larger negative bias in the number of events.

Evaluating the bias, MAE and RMSE of the forecasted vertical wind profiles at three of the sensor sites indicated that the experiments with turbine drag submodel produced the best forecasts in the lower layers (roughly 0-200 m). A subjective analysis of the forecasts for individual ramp event cases indicated that the main factor for the degradation in the ramp event performance score seen in the experiments that employed the turbine drag submodel was a frequent reduction of the amplitude of the event to just below the threshold required to qualify as a forecast of the event. Thus, the signal was typically present and the timing was often better in experiments with the turbine drag submodel but the amplitude had a significant negative bias.

A WRF configuration was selected as the baseline for subsequent experiments that produced power generation forecasts for the aggregate of TWRA wind generation facilities that had a 41.6% lower MAE and a 36.2% lower RMSE and also eliminated 88% of the bias relative to the baseline configuration over the 30-case experimental sample. It also produced a better simulation of the evolution (sequence and magnitude of the wind features at different sites) for most of the subjectively analyzed ramps.

This project also sought to improve the WRF model by using an improved method for data assimilation (DA), which is the process of incorporating observational data into the model. The research team tested three methods using the Gridpoint Statistical Interpolation (GSI) analysis system: an ensemble Kalman filter (ENKF), a three-dimensional variational method (3D-Var) and a hybrid method that incorporated elements of the previous two.

The various data assimilation methods were compared for a one-month sample of data as well as a set of 10 ramps. For forecasts of the mean wind speed, the DA methods produced mixed results, with the choice of best method depending on the location at which the forecasts were evaluated. For the ramp event forecasts, the hybrid method gave the best result, including the highest hit rate, the lowest false alarm ratio, the lowest bias, and the highest critical success index.

In the very short term (0-3 hours ahead), NWP models have limited utility due to the time required to collect and assimilate data and run the computational model. To address the need for accurate forecasts in the very short term, the team developed a forecasting system based on machine learning methods using data from the project sensors, wind generation sites, and weather monitoring stations in the TWRA with publicly available data. The forecasting system was developed with a 25-month data sample and evaluation over a 12-month period. The system was designed to produce 15-minute updates of three types of deterministic predictions: (1) a time series of the 15-minute average power production for the 12 intervals in the 0-3 hour look ahead period, (2) the maximum and minimum 60-minute ramp rate during the forecast period and (3) the occurrence/nonoccurrence of large ramps. Predictors were selected from a pool of 116 predictors via a screening algorithm.

Experiments were conducted to select the machine learning method, the values of the internal parameters of that method and the structure of the predictands. Based on these experiments, the Extreme Gradient Boosting (XGBoost) method was selected for this application. The optimal configuration of the forecast system for the time series prediction mode produced an average MAE of 4.1 % of capacity over the three-hour forecast period for the one-year evaluation period. The MAE ranged from 1.1% of capacity at 15 minutes to 7.2% at 180 minutes. The average three-hour reduction in MAE relative to a persistence forecast was 25.8% and ranged from 22.0% at 60 minutes to 29.8% at 180 minutes.

The best and final configuration for the 60-minute ramp rate prediction mode produced an MAE of 3.84% of capacity for the maximum ramp rate forecasts and 3.15% for the minimum ramp rate forecasts. This was an MAE reduction of 34.6% for the maximum ramp rate and 48.0% for the minimum ramp rate relative to a persistence forecast. The forecasts explained 36.6% of the variance in the maximum ramp rate and 50.7% of the variance of the minimum ramp rate.

For the large ramp event prediction mode, the best model configuration yielded a CSI of 27.8% for moderate-size upward ramps (300 MW in 60 minutes) and 7.1% for large ramps (800 MW in 60 minutes). The CSI scores for downward ramps were 35.7% for moderate events and 8.8% for large events.

All the improvements developed during this project were incorporated into the Improved Operational Forecast System (IOFS), which was evaluated in a six-month experiment that was designed to assess the performance of the IOFS in a multimethod ensemble approach generally considered the state-of-the-art method for operational power production forecasts. The baseline for the experiment was an ensemble of three NWP models operated by the NWS. A Model Output Statistics (MOS) procedure was applied to each model and the outputs were combined into an optimized ensemble. The research team created three more ensembles that consisted of the baseline ensemble and one additional member: the baseline operational forecast system (BOFS) used the WRF model configuration selected for this project in Task 2 without any additional data assimilation; the Enhanced Baseline Operational Forecast System (EBOFS) used the same WRF configuration and assimilated data from project sensors; and IOFS used the Task 2 WRF configuration, assimilated data from project sensors using the hybrid DA method of Task 4, and incorporated the very short-term statistical forecast (VSTF) system from Task 5. The same procedure of MOS and ensemble optimization were followed for the three project ensembles.

The baseline optimized NWS-MOS ensemble produced an MAE of 7.56% for the forecasts of the 15-minute average power production over the full 15-hour forecast period for the entire one-year evaluation period. The two initial project forecast systems (BOFS and EBOFS) added to the NWS ensemble did not improve upon the performance of the optimized NWS ensemble over the full forecast period. Due to the time and resource limitations of this project, the IOFS could only be evaluated over a six-month subsample of the one-year Forecast Improvement Assessment Experiment (FIAE) period. Over the six-month period, the IOFS ensemble reduced the 0-15 hour MAE of TWRA aggregate power generation forecasts from the baseline NWS ensemble by 6.7% without the VSTF component and 13.5% with the use of the VSTF. Most of the improvement was in the 0-3 hour portion of the 15-hour forecast window.

7.2 Impact of Project Sensors

Sensors used as part of this project contributed significantly to the forecast performance. For the very short-term forecasts using empirical methods, the project sensors reduced the mean absolute error (MAE) of the 15-minute average power production over the entire 0-3 hour forecast period by 7.3%. The reduction in MAE varied from about 2% at 15 minutes to nearly 10% at 180 minutes. The overall reduction in MAE (including data from nonproject sensors) compared to a persistence forecast ranged from about 22% at 60 minutes to nearly 30% at 180 minutes.

In forecasts of the maximum (upward) and minimum (downward) ramp rates, adding the project sensor data resulted in a 6.92% reduction in MAE and a 12.42% improvement in the forecast to observed correlation for upward ramps, and a 5.3% reduction in MAE and an 11.4% improvement in the forecast to observed correlation for downward ramps. Project data had the largest effect on ramp event forecasts for the largest events (800 MW upward ramp or 750 MW downward ramp in 60 minutes). For the largest upward ramps, the CSI increased from 2.9% to 7.1% when the project data were employed. The benefit was even larger for downward events, with the CSI increasing from 2.1% to 8.8% when predictors calculated from the project sensor data were added.

The sensor that had the largest impact on the very short-term forecasts was the radar wind profiler at Bena, which contributed 71% of the total reduction in MAE due to project sensors for the 15-minute average power production. The Bena wind profiler also had the largest impact on ramp rate forecasts, contributing about 80% of the total benefit from project sensors for upward ramps and about 60% for downward ramps.

The effect of project sensor data on the performance of the NWP models is more complex, and was not assessed in as great detail as for the machine learning methods. Comparing the performance of the BOFS with the EBOFS, which includes project sensor data, indicates the impact of that data on model performance. The MAE of the 15-minute average power production over the 0-15 hour forecast period was reduced by 1.8% for EBOFS compared to BOFS, while the mean bias over the period was reduced by 8.5%. However, when BOFS and EBOFS were run as parts of optimized ensembles, the MAE of the 15-minute average power production was slightly lower for BOFS (7.53% of capacity vs. 7.56% for EBOFS). The IOFS, which combined project data with WRF improvements and statistical forecasting, improved the optimized ensemble forecast MAE by 13.5% over the 0-15 hour forecast period compared with the NWS ensemble, with the greatest benefit seen in the first three hours. The concentration of the forecast accuracy improvement in the 0-3 hour period was most likely attributable to the fact that the reduction in forecast error was associated with the effective use of data from the project sensor network.

7.3 Project Benefits

The measurement component of this project produced a high-quality dataset of observations from a targeted sensor network.[2]

Data from project sensors had a significant impact on forecast skill in the very short term (0-3 hours ahead):

- 7.2% reduction in MAE of TWRA aggregate power production forecast
- 6.9% reduction in MAE of ramp rate forecast
- Increased CSI from 2.5% to 8.8% for large ramp events (>750 MW)
- Bena radar wind profiler having the largest impact on very short-term forecasts

This project produced several quantifiable improvements in wind speed forecasting that can be immediately implemented in forecasts provided to the California Independent System Operator, utilities and wind plant operators:

- Data assimilation improvements contributed to 6.7% reduction in MAE of TWRA aggregate power production
- Adding very short-term machine-learning forecasts contributed another 6.7% reduction in MAE of TWRA aggregate power production
- Overall, IOFS produced a 13.5% reduction in MAE of 0-15 hour forecasts of the 15-minute average TWRA power production relative to an optimized ensemble of forecasts from National Weather Service models.

7.4 Recommendations

This project suggests several areas for continuation and further research. Extending the IOFS evaluation period to a full year would improve the analysis and is fairly simple to achieve with the data already available. The performance of the IOFS at forecasting ramps could also be evaluated. Another possibility to leverage existing data would be the integration of observations and NWP outputs into a single training set for machine learning, which could potentially yield further reductions in forecast errors. There are also opportunities to continue optimizing the machine learning method by varying internal parameters or automating the predictor selection process.

An interesting area for more in-depth research is the role of wind turbine drag in atmospheric prediction models. This project found that including the turbine drag submodel in WRF led to significant decreases in the bias and MAE of mean wind speed forecasts, but degraded the performance of the ramp event forecasts. This problem could be addressed either from the perspective of improving NWP models to better capture the atmospheric processes near wind plants, or by altering the parameters for identifying ramp events in time series forecasts of wind speeds or power production.

[2] Available through Sonoma Technology, Inc. (http://www.sonomatech.com/).

This study benefitted from having data spanning more than two years from meteorological sensors installed for the purpose of improving wind ramp forecasts. Unfortunately, the sensors could not remain in place beyond the study period, so the benefits associated with those measurements are no longer available. Techniques such as the machine learning methods described in Chapter 5 perform better when given more data; for example, a longer training sample might have enabled the identification of seasonal trends among predictors. Additional measurements such as upper level winds and regional pressure differences could also provide valuable predictors. To produce a lasting improvement in wind ramp forecasts, there is a need for a stable network of targeted meteorological sensors to derive the greatest value from the data.

GLOSSARY

Metrics

Term	Definition
Bias	Average error over an evaluation sample
BR	Bias ratio: ratio of the number of predicted events to observed events
CSI	Critical success index: the ratio of successful forecasts of an event (hits) to the sum of the hits, misses (occurrence of an event with no forecast) and false alarms (forecasts of the event with no occurrence)
FAR	False alarm rate: ratio of forecasted events that did not occur (false alarms) to total number of forecasted events
HR	Hit rate: ratio of forecasted events that actually occurred (hits) to total number of observed events
MAE	Mean absolute error: average of the absolute deviations between the predicted and observed value of a forecast variable
MR	Miss ratio: ratio of observed events that were not forecasted (misses) to total number of observed events
RMSE	Root mean square error: the square root of the average of the squared deviations between the predicted and observed value of a forecast variable

Project-specific terms

Term	Definition
AMMT	Atmospheric Measurements and Modeling of the Tehachapi Wind Resource Area (project title)
BOFS	Baseline operational forecast system
EBOFS	Enhanced baseline operational forecast system
FIAE	Forecast improvement assessment experiment (Task 6)
IOFS	Improved operational forecast system
VSTF	Very short-term statistical forecast system (product of Task 5)

Other terms

Term	Definition
3D-Var	Three-dimensional variational assimilation: a data assimilation method
AGL	Above ground level
AWST	AWS Truepower, LLC (project partner)
California ISO	California Independent System Operator
DA	Data assimilation: incorporation of observations into a numerical model
EnKF	Ensemble Kalman filter: a recursive filter suitable for problems with a large number of variables, such as discretizations of partial differential equations in geophysical models
EnSRF	Ensemble square root filter
GBM	Gradient boosting machine: a machine learning method
GDAS	Global data assimilation system
GFS	Global forecast system: a numerical weather prediction model
GSI	Gridpoint statistical interpolation: a unified variational data assimilation system for both global and regional atmospheric model applications
LSM	Land surface model
MSL	Mean sea level
MOS	Model output statistics: a statistical procedure that uses predictors from the grid point output of an NWP model to train a statistical prediction model for a weather-dependent measured variable
NAM	North American mesoscale model
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction: the division of the US National Weather Service that executes daily cycles of NWP models for real-time forecast purposes
NOAA	National Oceanographic and Atmospheric Administration: the US Government agency that houses the National Weather Service
NWP	Numerical weather prediction: physics-based mathematical models of the atmosphere
O-A	Observation minus analysis
O-B	Observation minus background
PBL	Planetary boundary layer
RAP	Rapid Refresh Model: a physics-based atmospheric model that is operationally run by the US National Weather Service on an hourly cycle

Term	Definition
RASS	Radio acoustic sounding system
RWP	Radar wind profiler
SJV	San Joaquin Valley
STI	Sonoma Technology, Inc. (project partner)
TWRA	Tehachapi Wind Resource Area: the region of concentrated wind generation capacity encompassing Tehachapi Pass and adjacent regions of the Mojave Desert in southern California
UTC	Coordinated universal time
WRF	Weather research and forecasting: an open-source physics-based numerical weather prediction model
XGBoost	Extreme gradient boosting: a machine learning method

REFERENCES

Benjamin, S.G., S.S. Weygandt, J.M. Brown, M. Hu, C.R. Alexander, T.G. Smirnova, J.B. Olson, E.P. James, D.C. Dowell, G.A. Grell, H. Lin, S.E. Peckham, T.L. Smith, W.R. Moninger, and J.S. Kenyon (2016). "A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh." American Meteorological Society, *Mon. Wea. Rev*., **144**, 1669-1694.

California Energy Commission. (2017). *California Energy Almanac.* http://www.energy.ca.gov/almanac/

Chen, T. and C. Guestrin (2016). "XGBoost: A Scalable Tree Boosting System." KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pages 785-794. ISBN: 978-1-4503-4232-2

California And Nevada Smoke and Air Committee (CANSAC), 2015: http://cansac.dri.edu/cofframe.php?page=description.php.

Deppe, A.J., W.A. Gallus, and E.S. Takle (2013). "A WRF Ensemble for Improved Wind Speed Forecasts at Turbine Height." American Meteorological Society, *Weather Forecasting*, **28**, 212-228.

Fitch, A.C., J.B. Olson, J.K. Lundquist, J. Dudhia, A.K. Gupta, J. Michalakes, and I. Barstad (2012). "Local and Mesoscale Impacts of Wind Plants as Parameterized in a Mesoscale NWP Model." American Meteorological Society

Friedman, J. (1999). "Greedy Function Approximation: A Gradient Boosting Machine." Technical report, Stanford University. Available at https://statweb.stanford.edu/~jhf/ftp/trebst.pdf

Glahn, Harry R., and Dale A. Lowry (1972). "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting." *Journal of Applied Meteorology* **11**(8) pp.1203-1211.

Hu, X.M., J.W. Nielson-Gammon, and F. Zhang (2010). "Evaluation of Three Planetary Boundary Layer Schemes in the WRF Model." American Meteorological Society, *J. Appl. Meteor. Climatol*., **49**, 1831-1844.

Kleist, D.T., D.F. Parrish, J.C. Derber, R. Treadon, W.-S. Wu, and S. Lord (2009). "Introduction of the GSI into the NCEP Global Data Assimilation System." American Meteorological Society.

Kamisky, R., C.P. van Dam, C. MacDonald, and C. Collier (2016). "WindSense Project – Improvements in Short-term Wind Forecast for the Tehachapi Wind Resource Area Using a Targeted Observational Sensor Network," California Energy Commission, CEC-500-2016-070.

Manobianco, J., E.J. Natenberg, K. Waight, G. Van Knowe, J. Zack, D., Hanley, and C. Kamath (2011). "Limited Area Model-Based Data Impact Studies to Improve Short-Range Wind Power Forecasting," Lawrence Livermore National Laboratory, LLNL-ABS-491653.

Pan, Y., K. Zhu, M. Xue, X. Wang, M. Hu, S.G. Benjamin, S.S. Weygandt, J.S. Whitaker (2014). "A GSI-Based Coupled EnSRF-En3DVar Hybrid Data Assimilation System for the Operational Rapid Refresh Model: Tests at a Reduced Resolution." American Meteorological Society.

Rogers, R.E., A. Deng, D.R. Stauffer, B.J. Gaudet, Y. Jia, S-T. Soong, and S. Tanrikulu (2013). "Application of the Weather Research and Forecasting Model for Air Quality Modeling in the San Francisco Bay Area." American Meteorological Society, *J. Appl. Meteor. Climatol.*, **52**, 1953-1973.

Skamarock,W., J.B. Klemp, J. Dudhia, D.O. Gill, D.M. Barker, M.G. Duda, X.-Y. Huang, W. Wang, and J.G. Powers (2008). "A Description of the Advanced Research WRF Version 3." NCAR Technical Note NCAR/TN-475+STR, doi:10.5065/D68S4MVH.

Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467pp.

Wu, W-S. (2005). "Background Error for NCEP's GSI Analysis in Regional Mode." World Meteorological Organization.

Zack, J., E.J. Natenberg, S. Young, J. Manobianco, and C. Kamath (2010). "Application of Ensemble Sensitivity Analysis to Observational Targeting for Short Term Wind Speed Forecasting." LLNL Technical Report LLNL-TR-424442.

APPENDICES

The following appendices are available as a separate publication (CEC-500-2018-002-AP-A-F):

APPENDIX A - Task 2: Model Sensitivity Experiments

APPENDIX B - Task 3: Field Measurements

APPENDIX C - Task 4: Short Term wind Ramp Forecasting Improvement

APPENDIX D - Task 5: Very Short-Term Statistical Wind Power Forecast Tool

APPENDIX E - Task 6: Wind Ramp Forecast System Evaluation and Finalization

APPENDIX F - Task 7: Model Sensitivity Experiments