



The Verification and Validation Strategy Within the Second Wind Forecast Improvement Project (WFIP 2)

Caroline Draxl,¹ L. K. Berg,² L. Bianco,³ T. A. Bonin,^{3,8} A. Choukulkar,^{3,8} A. Clifton,⁴ J. W. Cline,⁵ I. V. Djalalova,^{3,8} V. Ghate,⁷ E. P. Gritmit,⁶ K. Holub,³ J. S. Kenyon,^{3,8} K. Lantz,^{3,8} C. Long,³ J. K. Lundquist,^{1,9} J. McCaa,⁶ K. McCaffrey,^{3,8} J. F. Newman,¹⁰ J. B. Olson,^{3,8} Y. Pichugina,^{3,8} J. Sharp,¹¹ W. J. Shaw,² N. H. Smith,¹² and M. D. Toy^{3,8}

1 National Renewable Energy Laboratory

2 Pacific Northwest National Laboratory

3 National Oceanic and Atmospheric Administration, Earth System Research Laboratory

4 WindforS, University of Stuttgart

5 U.S. Department of Energy

6 Vaisala

7 Argonne National Laboratory

8 University of Colorado, Cooperative Institute for Research in Environmental Sciences

9 University of Colorado, Department of Atmospheric and Oceanic Sciences

10 RESurety, Inc.

11 Sharply Focused, LLC

12 Texas Tech University

**NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC**

**Technical Report
NREL/TP-5000-72553
November 2019**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308



The Verification and Validation Strategy Within the Second Wind Forecast Improvement Project (WFIP 2)

Caroline Draxl,¹ L. K. Berg,² L. Bianco,³ T. A. Bonin,^{3,8} A. Choukulkar,^{3,8} A. Clifton,⁴ J. W. Cline,⁵ I. V. Djalalova,^{3,8} V. Ghate,⁷ E. P. Gritmit,⁶ K. Holub,³ J. S. Kenyon,^{3,8} K. Lantz,^{3,8} C. Long,³ J. K. Lundquist,^{1,9} J. McCaa,⁶ K. McCaffrey,^{3,8} J. F. Newman,¹⁰ J. B. Olson,^{3,8} Y. Pichugina,^{3,8} J. Sharp,¹¹ W. J. Shaw,² N. H. Smith,¹² and M. D. Toy^{3,8}

1 National Renewable Energy Laboratory

2 Pacific Northwest National Laboratory

3 National Oceanic and Atmospheric Administration, Earth System Research Laboratory

4 WindforS, University of Stuttgart

5 U.S. Department of Energy

6 Vaisala

7 Argonne National Laboratory

8 University of Colorado, Cooperative Institute for Research in Environmental Sciences

9 University of Colorado, Department of Atmospheric and Oceanic Sciences

10 RESurety, Inc.

11 Sharply Focused, LLC

12 Texas Tech University

Suggested Citation

Draxl, C., L. K. Berg, L. Bianco, T. A. Bonin, A. Choukulkar, A. Clifton, J. W. Cline, et al. 2019. *The Verification and Validation Strategy Within the Second Wind Forecast Improvement Project (WFIP 2)*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-5000-72553. <https://www.nrel.gov/docs/fy20osti/72553.pdf>.

**NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-5000-72553
November 2019

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored [in part] by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

List of Acronyms

AMS	American Meteorological Society
BPA	Bonneville Power Administration
CBS	cross-barrier synoptic
CBT	cross-barrier thermal
CO	convective outflow
DAP	Data Archive and Portal
DOE	U.S. Department of Energy
EERE	Office of Energy Efficiency and Renewable Energy
EF	easterly flow
EMAT	Experiment to Model Analysis Table
GUI	graphical user interface
HRRR	High-Resolution Rapid Refresh
MAE	mean absolute error
MW	mountain waves
NCAR	National Center for Atmospheric Research
NCEP	please provide
NOAA	National Oceanic and Atmospheric Administration
NOAA ESRL	NOAA Earth System Research Laboratory
NWP	numerical weather prediction
OT	other, light and variable
PIRT	Phenomena Identification and Ranking Table
RAP	Rapid Refresh
RMSE	root-mean-square error
RT&M	Ramp Tool and Metric
SL	stable layer
TW	topographic wakes
UTC	Universal Time Coordinated
V&V	verification and validation
WFIP 2	Second Wind Forecast Improvement Project
WL	window length
WRF	Weather Research and Forecasting

Table of Contents

List of Acronyms	iii
Table of Contents	iv
List of Figures	v
List of Tables	v
1 Introduction	1
2 Definition of V&V	2
2.1 Verification.....	2
2.2 Validation.....	2
2.3 V&V in WFIP 2	2
2.4 Ad-Hoc Validation Versus Formal Validation.....	3
3 V&V Tools	4
3.1 The Experiment To Model Analysis Table	4
3.2 WFIP 2 Weather Taxonomy.....	4
3.3 WFIP 2 Event Log.....	5
3.4 Case Study Report Template.....	5
3.5 Communication: Interviews and Meetings.....	5
3.6 Data Archive and Portal	6
3.7 Common Metrics.....	6
3.8 Common Case Study Data Set To Test Code.....	7
4 Results	8
4.1 Evaluation of the Operational HRRR-NCEP for Various Event Types.....	8
4.2 Seasonal Evaluation of MAE	9
4.3 Ramp Events	10
4.4 Wind Power Forecasts.....	12
4.5 Conference Contributions	14
5 Summary	17
5.1 Lessons Learned.....	17
5.2 Recommendations for the Future	18
References	19

List of Figures

Figure 1. Screenshot of parts of the WFIP 2 EMAT.....	4
Figure 2. Schematics and pictures for each of the five major categories describing weather phenomena in the Columbia River Gorge	5
Figure 3. Adopting the WFIP 2 V&V culture requires several aspects which are listed in this figure. At the core of the V&V strategy is the team. This figure lists the most important aspects of the WFIP 2 V&V strategy.....	6
Figure 4. Model bias in rotor-layer (50 – 150 m) wind speed at two sites, (top) Wasco and (bottom) Arlington, as a function of the different event types (CBS: Cross-Barrier Synoptic, CBT: Cross-Barrier Thermal, SL: Stable Layer, MW: Mountain Waves, TW: Topographic Wakes, CO: Convective Outflow, EF: Easterly Flow, OT: other, light and variable). The + indicates that the event is building, and the - indicates that the event is decaying. The bias was calculated from scanning lidars and the 3-km HRRR model simulation at forecast hour 3. ...	9
Figure 5. Diurnal composite time-height sections depicting the change in MAE of wind-speed forecasts (color fills, m s^{-1}) between the experimental and control HRRR configurations for the (a) spring, (b) fall, (c) summer, and (d) winter reforecast periods as verified against 915-megahertz (MHz) radar wind profilers. Shades of blue (red) correspond to MAE improvements (degradations) in the experimental configuration over the control configuration. (source: Figure 6 in Olson et al. 2019)	10
Figure 7. Schematic that shows how the model simulations were stitched together	11
Figure 8. a) The ramp skill score for each model seasonally and annually with error bars indicating statistical significance ($\sigma/(N-1)^{1/2}$) and (b) the percentage improvement in the forecast skill due to improved physics for the 3-km HRRR (blue) and 750-m HRRRNEST (red) models in the bottom panel (source: Figure 14 in Olson et al. 2019).....	12
Figure 9. Normalized RMSE (m/s) for wind power forecasts for the BPA fleet aggregate using the experimental (blue) and control (red) 750-m HRRR simulations for the months of January 2017, April 2016, July 2016, and October 2016, grouped by forecast hours 1–6, averaged from 7–15, and averaged 16–24 as indicated on the x-axis	13
Figure 10. Schematic of error reductions. Improved simulations exhibit lower errors than baseline NWP output. Postprocessing the baseline NWP output reduces the error considerably. Postprocessing improved simulations yields the lowest errors.....	13

List of Tables

Table 1. WFIP 2 Related Talks and Poster Presentations at the 98 th AMS Annual Meeting in Austin, Texas	14
--	----

1 Introduction

The Second Wind Forecast Improvement Project (WFIP 2), in the complex terrain of the Columbia Gorge, focuses on a set of weather phenomena that poses particular challenges for wind and wind power forecasting. It further aims to understand and improve the skill (i.e., accuracy) of weather forecast models, particularly the National Oceanic and Atmospheric Administration's (NOAA's) High Resolution Rapid Refresh (HRRR) model, in complex terrain. The project also includes an extensive field campaign in the Columbia River Gorge, during which data from many different instruments were collected. The project is described in detail in three journal papers in the Bulletin of the American Meteorological Society (Shaw et al. 2019; Wilczak et al. 2019; Olson et al. 2019).

WFIP 2 is funded by the U.S. Department of Energy (DOE). Scientists from four U.S. national laboratories, NOAA, Vaisala, and universities contribute to the project with wide and varied interests and skill sets. Therefore, coordinated verification and validation (V&V) efforts across these member institutions allows for the development of a clear picture of model improvements and scientific findings within WFIP 2. The WFIP 2 V&V team is tasked with providing tools, methods, and guidance to enable repeatable, metrics-based assessment of the Weather Research and Forecasting (WRF) model and associated modeling suites for analysis and forecasting of mesoscale weather phenomena that are important for wind energy in the Columbia River Gorge and other parts of the continental United States.

This report summarizes what has been accomplished within the duration of the WFIP 2 project (4 years) by the V&V team. It is intended to be used to look up expectations, definitions, tools, results, and agreed-upon processes. Lessons learned toward an integrated V&V process across DOE's Atmosphere to Electrons (A2e) initiative are presented as well.

2 Definition of V&V

As a project funded by the DOE's Office of Energy Efficiency and Renewable Energy (EERE), WFIP 2 follows the DOE EERE definitions of verification and validation. These are engineering definitions that can be found in, e.g., the American Institute of Aeronautics and Astronautics' Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (<http://arc.aiaa.org/doi/abs/10.2514/4.472855.001>).

2.1 Verification

Verification is concerned with checking the mechanics of the software code rather than checking that the model's physics are correct. Because the code mechanics may include the use of discrete equations to represent processes that were defined using partial differential equations, there is the risk that numerical errors may be introduced and verification seeks to ensure that these errors have been identified and minimized. A model might also include some kind of iterative process to find a solution, which also requires checking.

2.2 Validation

Validation is determining the degree to which the model represents the real world for a particular application. Validation should be carried out after verification to ensure that the validation process identifies errors in the model physics. Validation of an unverified model may otherwise just identify numerical or coding problems.

2.3 V&V in WFIP 2

WFIP 2 brings together many organizations that have wide and varied interests and skill sets. Coordinated V&V across these groups allows them to build on the cumulative WFIP 2 effort and allows the team to develop a clear picture of model improvement in WFIP 2.

The model of interest in WFIP 2 is the High-Resolution Rapid Refresh (HRRR), which is based on the Weather Research and Forecasting (WRF) model. During WFIP 2, members of the WFIP 2 team made improvements to various parts of the WRF model physics that are targeted at specific pre-existing problems in WRF that have a large impact of forecasts of the wind energy resource. These problems have been identified by the WFIP 2 team or elsewhere in the community.

In WFIP 2, V&V is a process that starts by engaging stakeholders to identify and document their needs to develop a validation strategy. To increase the impact of the validation process, results are then reported back to stakeholders in ways that are useful to them. Therefore, the mission of the WFIP 2 V&V effort is to document and communicate model performance improvements resulting from the WFIP 2 effort in a way that is meaningful to stakeholders, including the weather modeling community, the wind energy industry, and national funding agencies. Some stakeholders might be interested in parameters within WRF related to turbulence, for example, while others in the wind-energy community might be more interested in the ability of the model to predict ramps in hub-height wind speeds. It is important to note that stakeholders will not provide consensus; rather, they will provide insight.

In a formal validation process, metrics for comparing the model output and field observations are agreed upon with stakeholders who also rank these comparisons in terms of importance. The metrics are related to a specific phenomenon. For example, wind speed ramps are a phenomenon of high importance to multiple stakeholders.

HRRR model verification in WFIP 2 was assumed to have mainly been carried out by the National Center for Atmospheric Research (NCAR), which maintains the WRF code, and included single-column model simulations. NOAA also performed regression tests and simple code tests for the HRRR for verification.

The goal of the WFIP 2 V&V effort is to provide tools (Section 3), methods, and guidance to enable repeatable, metrics-based assessment of WRF and associated modeling suites for analysis and forecasting of mesoscale weather phenomena that are important for wind energy in the Columbia River Gorge and the continental United States.

2.4 Ad-Hoc Validation Versus Formal Validation

Many institutions deployed instruments in the field in WFIP 2 and are interested in how well those devices worked. An easy way to do that is to compare those data to model results. On the other hand, model developers want to test their improvements quickly for certain critical cases. They then compare model simulations and observations at a few sites qualitatively. We call this *ad-hoc validation*, for which a few instruments are usually used for a short period. Ad-hoc validation can be very helpful but is more powerful when done in a way that ties in to the rest of the WFIP 2 V&V activities. The challenge with ad-hoc validations is that they are often done and presented differently, using different assumptions, which makes it difficult to see what was learned and how broadly applicable that was. Therefore, we have been carrying out interviews to bring the data and lessons learned from these validations together in one place. This allows us to display all of the information together coherently. We are calling this the *survey of ad-hoc validations*. In addition, NOAA maintained a real-time model-observation evaluation website (http://wfip.esrl.noaa.gov/psd/programs/WFIP_2/, last accessed September 26, 2018), that was used for daily monitoring of model forecasts during the duration of the field campaign (Wilczak et al. 2019).

Formal, or quantitative, validation is a bigger effort where a few or hundreds of different observations and parameters are compared to a model output using formal agreed-upon metrics. This validation is intended to get information about the model performance. In WFIP 2, once all the data from the instruments and models are collected, we can start by selecting the case studies from an event log that can be analyzed in a systematic way. *Case studies* are 1–3-day periods where one type of weather regime dominates, or something interesting is happening that challenges the model. In these case study periods, observations will be compared to the model outputs using metrics set out in the developed Experiment to Model Analysis Table (EMAT). We are also running a retrospective simulation of several months to test the model across the entire contiguous United States. This is to make sure we improve the performance in the Columbia River Gorge as well as other locations across the United States.

3 V&V Tools

3.1 The Experiment To Model Analysis Table

The Phenomena Identification and Ranking Table (PIRT) is one of the most common tools in a formal V&V process. The PIRT sets out the what and why of a validation process. It captures information about each stakeholder’s priorities. Given the large number of participants in WFIP 2, instead of a PIRT, we developed an Experiment to Model Analysis Table (Figure 1), which served as a PIRT and was designed to be a tracking and management tool that brings together the science questions that WFIP 2 wants to address with stakeholder interests, weather regimes, and instruments deployed in the field. As a living document, it also helps keep an overview of who is leading what work. It was first developed by modelers, reviewed by instrument owners and industry participants, and iterated to get all instruments included.

SUMMARY		Description of Experiment		User should enter a short description of the investigation in this row ->	ABL depth	80-m wind speed errors	80-m wind speed ramp metric	Turbulence dissipation rate	Mesoscale pressure gradients	
Summary of this specific investigation; auto populated except for the description		Justification	Why (Row 11)	Checks if this is defined	Complete	Complete	Complete	Complete	Complete	
		Experiment definition	What (Row 43)	Checks if this is defined	Complete	Complete	Complete	Complete	Started	
		Contribution to WFIP	How (Row 49)	Checks if this is defined	Complete	Complete	Complete	Complete	Complete	
		Who (Row 55)		Checks if this is defined	Complete	Complete	Complete	Complete	Complete	
		Priority	When (Row 61)	Checks if this is defined	Complete	Complete	Complete	Complete	Not Started	
		Notes	Huh (Row 63)	Checks if this is defined	Complete	Complete	Complete	Complete	Not Started	
		Instruments to be used	Where (Row 66)	Checks if this is defined	Complete	Complete	Complete	Complete	Complete	
					count					
WHY?	What is the motivation for this measurements or science question that is being addressed? Select either a weather regime, or one of the ISP sections	Weather Regimes	Cross-barrier thermal	CBT	Used in event log	6		y		
			Cross-barrier synoptic	CBS+	Used in event log	6		y		
				CBS	Used in event log	6		y		
				CBS-	Used in event log	6		y		
			Stable Layer	SL+	Used in event log	9		y		y
				SL	Used in event log	9		y		y
				SL-W	Used in event log	10		y		y
				SL-C	Used in event log	8		y		y
				SL-R	Used in event log	8		y		y
			Mountain Waves	MW	Used in event log	6		y		
			Topographic Waves	TW	Used in event log	10		y		y
			Convective Outflows	CO	Used in event log	6		y		
			Easterly Flow (no SW flow EF)	EF	Used in event log	6		y		
			Other (light and variable)	OT	Used in event log	3		y		
			https://docs.g.../?usp=sharing			ISP section 3.1	1	y		
			 vertical wind shear in generating			ISP section 3.2	0			
		Analysis	Evaluate the role of the surface energy budget on hub-height flow characteristics in varying terrain			ISP section 3.3	1			
			Flow/regime characterization			ISP section 3.4	0			
			Case study analysis to understand physics			ISP section 3.5	0			
			Large forecast error analysis			ISP section 3.6	0			
Dissipation rate analysis				ISP section 3.7	1			y		
Microbarograph analysis				ISP section 3.8	1					
Evaluate the role of the surface radiation budget, aerosols, and clouds in different meteorological regimes				ISP section 3.9	1			y		
Evaluate usefulness of instrument array for real time forecasting				ISP section 3.10	1		y			
ISP: Improvement of Numerical Techniques	Model improvements for high-resolution mesoscale simulations over complex terrain				ISP section 4	2		y		
ISP: Improvement of Model Physics	Evaluation and development of surface layer models for complex terrain				ISP section 5.1.1	0				
	Improvement of bottom boundary conditions for complex terrain			ISP section 5.1.2	0					
	Model development of scale appropriate surface layer models for complex terrain			ISP section 5.1.3	0					
	Evaluation and improvement of PBL parameterizations			ISP section 5.2	1					
	Data Assimilation Research			ISP section 5.3	1					
	Evaluation of wind farm wake parameterization			ISP section 5.4	1					
	Ensemble Analysis of model physics			ISP section 5.5	1		y			

Figure 1. Screenshot of parts of the WFIP 2 EMAT

3.2 WFIP 2 Weather Taxonomy

Meteorological phenomena observed in the study area were grouped into five major categories (Figure 2): gap flows, cold pool mix-outs, mountain waves, orographic waves, and marine pushes or thermal troughs. Furthermore, for these events we distinguish the formation, decay,

and persistence of the event. These classifications were used in the Event Log (Section 3.3), in the publications, and when referring to events in general.

3.3 WFIP 2 Event Log

For several months, the weather in the Columbia River Gorge was reviewed in weekly weather discussions set up for that purpose and subsequently logged using the weather taxonomy (Section 3.2). The weather, model performance, the importance of the weather for the wind energy industry, as well as a description of the weather, were documented. Through regular exposure to the peculiarities of the weather in the study area, as well as model performance, the team could become familiar with the Columbia River Gorge's weather and learn about shortcomings of the HRRR right from the beginning of the project. The Event Log is available to identify case studies.

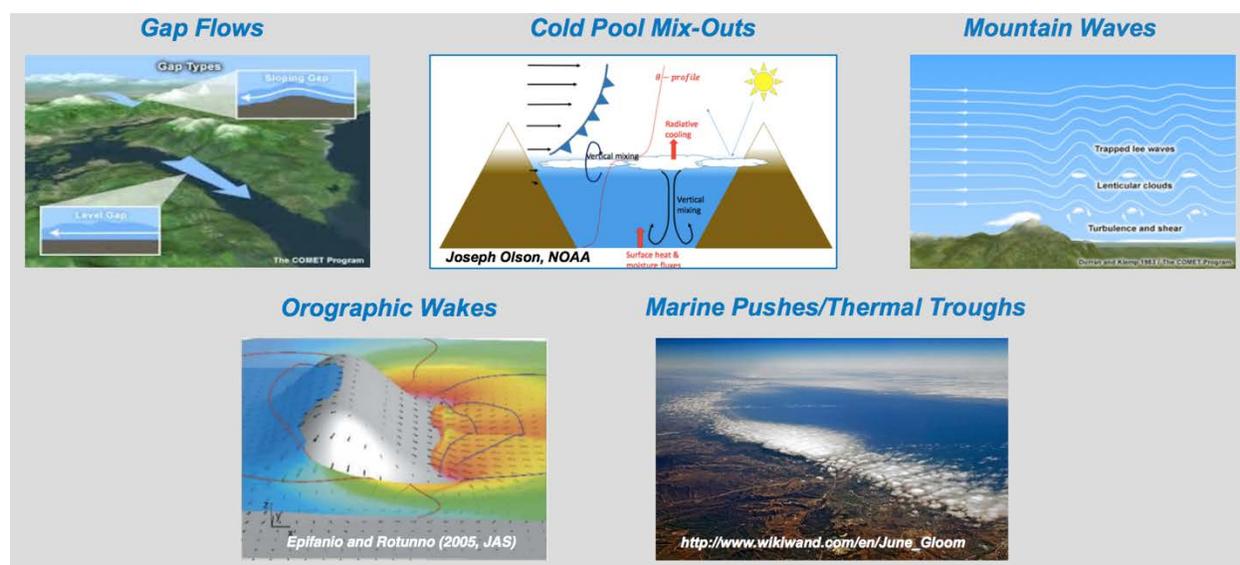


Figure 2. Schematics and pictures for each of the five major categories describing weather phenomena in the Columbia River Gorge

3.4 Case Study Report Template

For standardized reporting of case studies, a case study report template was developed that follows a paper outline and includes free-form results and explanations. The intent was to standardize results with the aim to draw conclusions more easily at the end of the project. The setup of the template was chosen to not duplicate efforts for team members who were going to write journal articles.

3.5 Communication: Interviews and Meetings

Regular communication between the team members turned out to be one of the most important tools. In fact, the team was considered the core of the V&V culture (Figure 3). Adopting a new culture is never easy (Oberkampf and Roy 2012) and adopting a culture has a lot to do with different personalities. Sitting together in a room and discussing who is doing what, discussing results, and sharing experiences was extremely helpful. Toward the end of the project, a workshop was organized to compare results and to make sure everyone's code was bug free.

Workshops and regular meetings also made sure no one was stepping on anyone's toes and that work was optimized and not duplicated. To that end, interviews were conducted at earlier stages of the project to list who was working on what, giving team members the opportunity to join anyone's particular effort.

3.6 Data Archive and Portal

The Data Archive and Portal (DAP) (<https://a2e.energy.gov/about/dap>) is a publicly available interface where data from several projects across DOE are collected. For WFIP 2, it provides access to simulations, observations, and metadata for specific case studies or the whole project period, respectively.

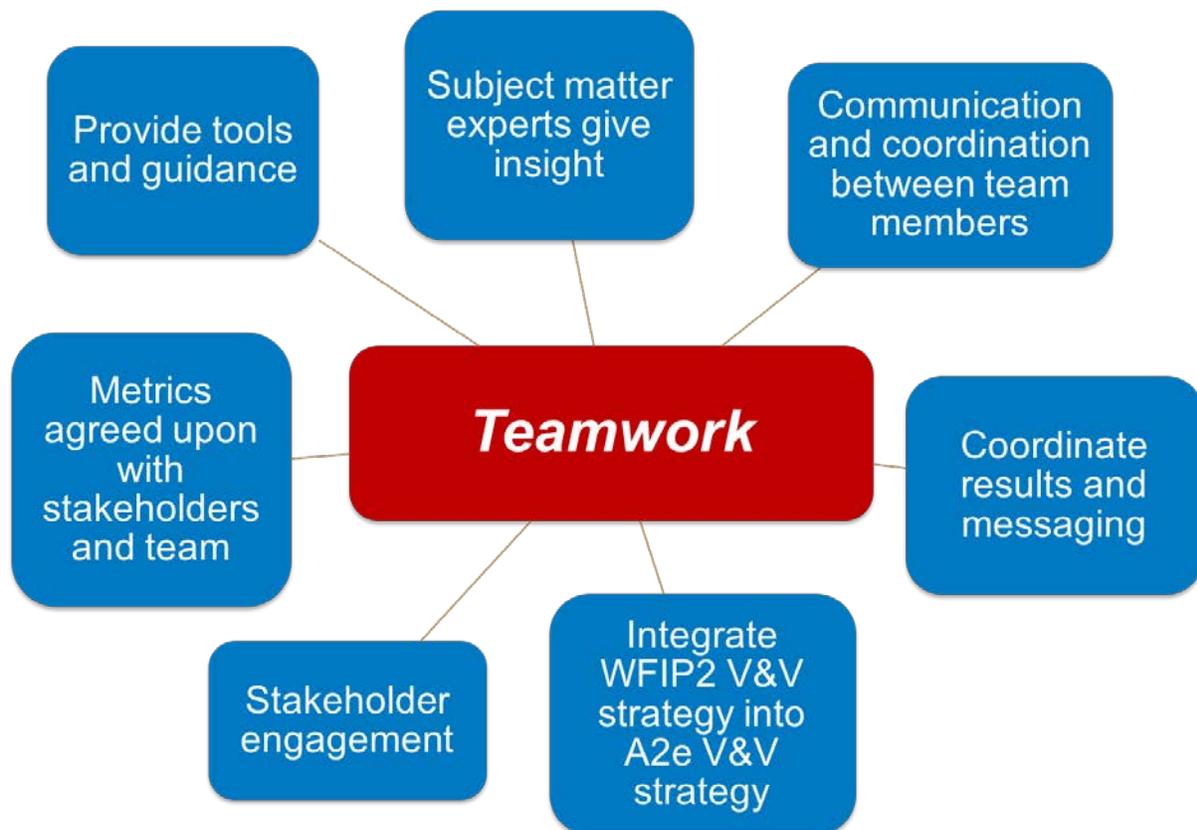


Figure 3. Adopting the WFIP 2 V&V culture requires several aspects which are listed in this figure. At the core of the V&V strategy is the team. This figure lists the most important aspects of the WFIP 2 V&V strategy.

3.7 Common Metrics

The key meteorological variable for WFIP 2 is the 80-meter (m) or 100-m wind speed in areas where there are wind power plants. The wind ramp metric, a metric to determine whether rapid wind changes were captured by the model, is the most important metric because it can be compared to the results from WFIP1. The Ramp Tool can be found on the NOAA Earth System Research Laboratory's (ESRL's) website (https://www.esrl.noaa.gov/psd/products/ramp_tool/)

and is available as either an executable file or a set of Matlab user files. In either case, the tool is run through a graphical user interface (GUI) in which you select from several options to customize it to your needs. The tool is described in an open-access article in the American Meteorological Society's Weather and Forecasting journal (<http://journals.ametsoc.org/doi/pdf/10.1175/WAF-D-15-0144.1>). Other important metrics are the mean absolute error (MAE), root-mean-square error (RMSE), and bias of bulk rotor layer statistics and wind power.

However, wind speed is only one output from the numerical weather prediction (NWP) modeling suite and so does not provide the information required to understand or detect model errors. Therefore, other variables should be analyzed to see that an improved model is “getting things right for the right reason.” Any new metrics needed for these evaluations should be carefully documented.

3.8 Common Case Study Data Set To Test Code

The team selected to rely on a common set of test data for validation to support multiple validation analysis codes. Therefore, we created a common test data set upon which the team could test their code. The idea was that if a code would yield the same metrics as the available plots and results from the test data set, it was assumed that the code was bug free. This approach assured that each team member could use the programming language and code they preferred.

4 Results

The hourly-updating 13-kilometer (km) Rapid Refresh (RAP) and the 3-km HRRR numerical weather prediction models were targeted for improvement. A 750-m HRRR nest around the field study area was also run. The WFIP 2 field observations were assimilated into the operational RAP and HRRR during the field campaign. The initial conditions for the case studies and year-long reforecasts were initialized off the operational RAP. In contrast, the 10-day retrospective periods were run in fully cycled forecast system mode (with data assimilation) for both the RAP and HRRR: The RAP was cycled (restarted) every hour, but the HRRR was only cycled every third hour. In the 750-m nest no additional data were assimilated, but that domain was initialized off of the 3-hour (hr) HRRR forecast and run concurrently within the HRRR out through hour 24. Model validation against multiple types of measurements was performed.

Model changes were tested in:

- Day-long case studies (only HRRR and its 750-m nested domain [HRRRNEST])
- Ten-day retrospective periods (RAP, HRRR, and 750-m HRRRNEST)
- Year-long reforecasts (only HRRR and 750-m HRRRNEST)

where both control and experimental configurations were simulated.

The results were published in the Bulletin of the American Meteorological Society (e.g., Olson et al. 2019 and Wilczak et al. 2019) and other journals (e.g., Pichugina et al. 2019). Additionally, several other journal articles are envisioned to be submitted in the coming months and years. All of the case studies that have been worked on have been summarized in a common case study document, which was a deliverable to DOE from Vaisala. Here we show validation examples.

4.1 Evaluation of the Operational HRRR-NCEP for Various Event Types

The operational HRRR from the National Centers for Environmental Prediction (NCEP) was evaluated during the whole WFIP 2 period. The ability of the model to simulate winds, boundary layer heights, temperature profiles, and radiation at the surface can be investigated at various observation sites for the different event categories. Results from this project may provide some quantitative guidance on which events and variables are the best and most poorly simulated, which could be useful to both forecasters (when to trust the model) and modelers (which areas need improvement). Figure 4 shows model bias in wind speed at two sites, Wasco and Arlington, as a function of the different event types.

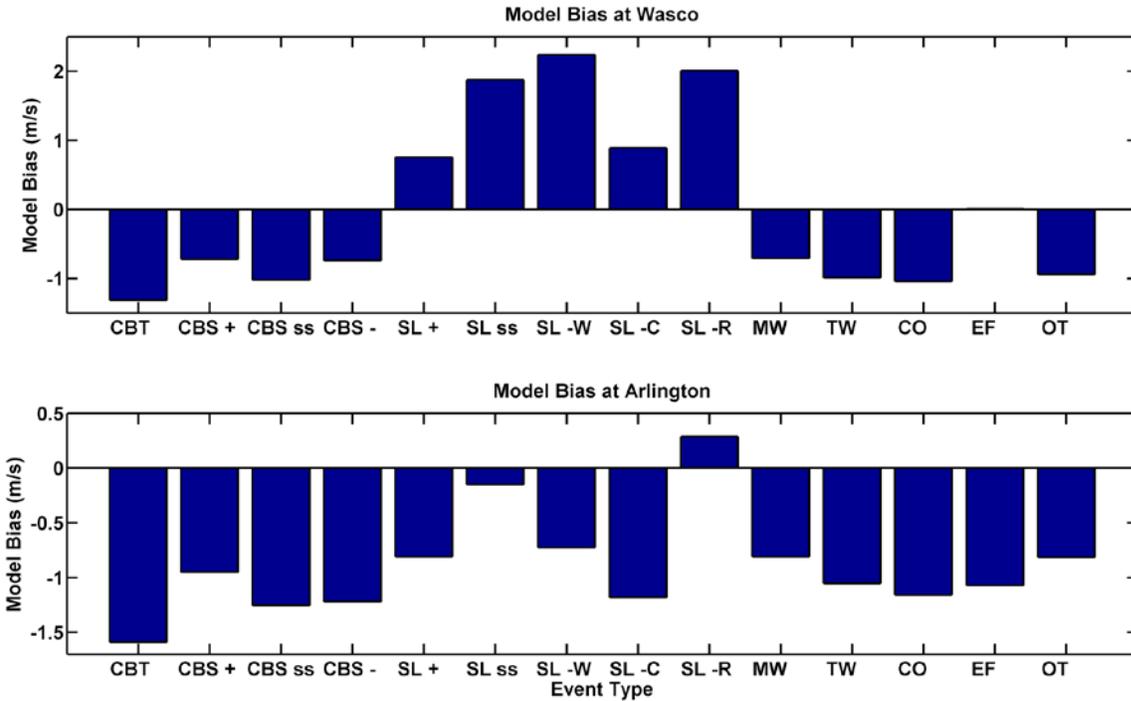


Figure 4. Model bias in rotor-layer (50 – 150 m) wind speed at two sites, (top) Wasco and (bottom) Arlington, as a function of the different event types (CBS: Cross-Barrier Synoptic, CBT: Cross-Barrier Thermal, SL: Stable Layer, MW: Mountain Waves, TW: Topographic Wakes, CO: Convective Outflow, EF: Easterly Flow, OT: other, light and variable). The + indicates that the event is building, and the - indicates that the event is decaying. The bias was calculated from scanning lidars and the 3-km HRRR model simulation at forecast hour 3.

4.2 Seasonal Evaluation of MAE

The MAE of wind-speed forecasts in the rotor layer and above between the improved and control HRRR configurations for the four seasons (spring, summer, fall, and winter) are shown in Figure 5. In Figure 5, blue indicates improvements and red indicates degradations, showing that rotor-layer winds were mostly improved but model simulations were degraded in the spring and summer months between 19–02 Universal Time Coordinated (UTC) (i.e., 11–18 local time).

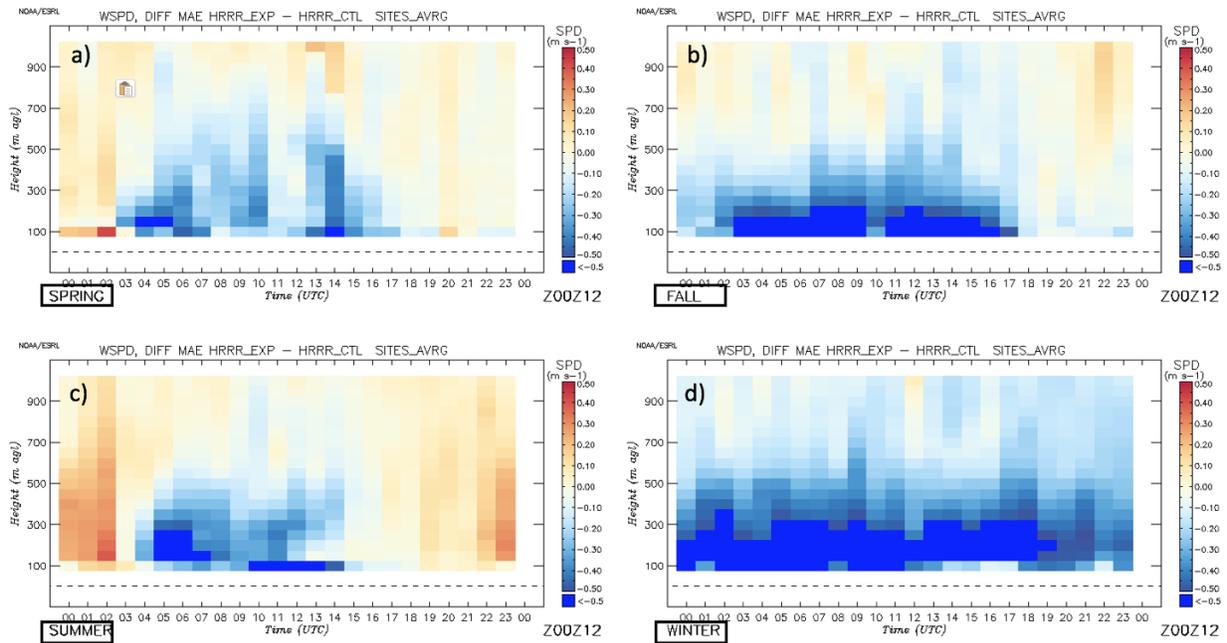


Figure 5. Diurnal composite time-height sections depicting the change in MAE of wind-speed forecasts (color fills, m s^{-1}) between the experimental and control HRRR configurations for the (a) spring, (b) fall, (c) summer, and (d) winter reforecast periods as verified against 915-megahertz (MHz) radar wind profilers. Shades of blue (red) correspond to MAE improvements (degradations) in the experimental configuration over the control configuration. (source: Figure 6 in Olson et al. 2019)

4.3 Ramp Events

Ramp events are important for wind operators because they need to keep the system in balance between conventional and renewable energy generation. The Ramp Tool and Metric (RT&M) is used to calculate the skill of the HRRR and the HRRRNEST in control and experimental mode at forecasting 80-m wind speed ramp events at 19 sodar locations during four reforecast periods (April, July, October 2016, and January 2017). The skill scores are computed for 20 different power ramp definitions (Figure 6 of Bianco et al. 2016) then averaged over the entire matrix of ramp definitions (Figure 6) to produce a single skill value.

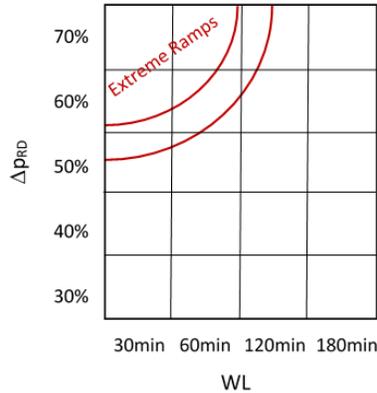


Figure 6. Schematic of a ramp matrix, with window length (WL) and power thresholds as in Bianco et al., 2016. Extreme ramps are in the top-left corner, and low-amplitude ramps of longer duration are in the bottom-right corner.

Because the simulations for the reforecast runs were cycled every 12 hours (at Z00 and Z120) to create long time series over which to run the RT&M, the model simulations had to be stitched together, as shown in Figure 7.

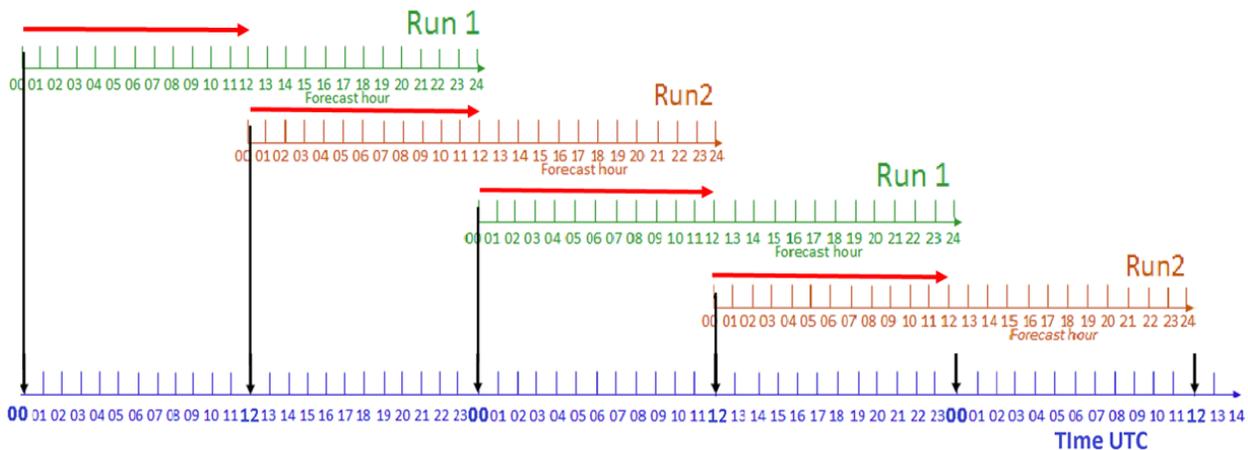


Figure 7. Schematic that shows how the model simulations were stitched together

After the time series of 80 m wind speed of model output and sodar observations were created, they were converted to the idealized power at each sodar location and then averaged to create the aggregated time-series of idealized power over the entire region. The RT&M was then run to compute the skills of the models. The ramp skill improvement is clearly visible (Figure 8) for the 3-km HRRR domain in the spring and winter with the largest improvement of nearly 60% in winter, being statistically significant as seen by the non-overlapping error bars in the upper panel between the control and experimental HRRR models. Conversely, the simulations were degraded in summer and fall. The ramp simulations of the 750-m domain show a degradation in winter only. On average, both simulations show less than 5% improvement over the entire year. %-improvement was calculated using the full set of 80-m wind speeds from 19 sodars.

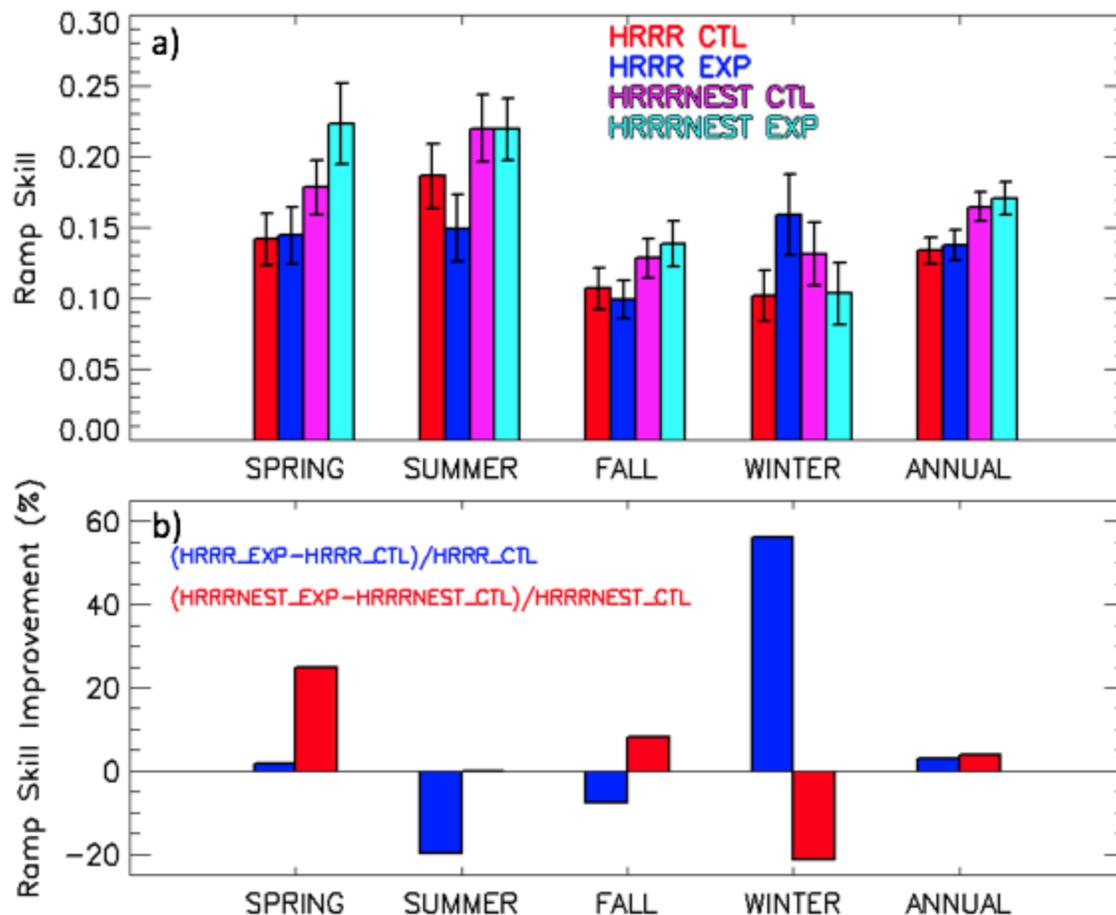


Figure 8. a) The ramp skill score for each model seasonally and annually with error bars indicating statistical significance ($\sigma/(N-1)^{1/2}$) and **(b)** the percentage improvement in the forecast skill due to improved physics for the 3-km HRRR (blue) and 750-m HRRRNEST (red) models in the bottom panel (source: Figure 14 in Olson et al. 2019)

4.4 Wind Power Forecasts

Vaisala calculated error metrics for statistically postprocessed wind power forecasts using the 750-m HRRRNEST reforecast data set. Figure 9 shows the Bonneville Power Administration (BPA) fleet aggregate normalized RMSE as a function of the forecast horizon for the experimental (blue) and control (red) simulations during the months of January 2017, April 2016, July 2016, and October 2016, representing winter, spring, summer, and fall results. In general, small improvements are seen in winter and fall but some degradations are noticeable in spring and summer. These results are not completely consistent with the wind ramp analysis shown in Figure 8 where the largest improvements in the 750-m HRRR domain are seen during the spring. However, more significant improvements are available in the 3-km HRRR domain, as shown in Figure 5, for raw model output in the turbine rotor layer. This is because at 750 m, not all improvements implemented into the model code are activated; some of these processes are not parameterized at that scale. However, as a general guide, the 750-m simulations are more accurate than the 3-km simulations even though the relative improvement from control to experimental simulations is decreased.

Raw model improvements are larger than those from the wind power forecasts shown here because these forecasts were statistically postprocessed (Figure 9). This means that because the forecasts were statistically corrected to past observations, this corrects for some of the deficiency that was present in the control runs.

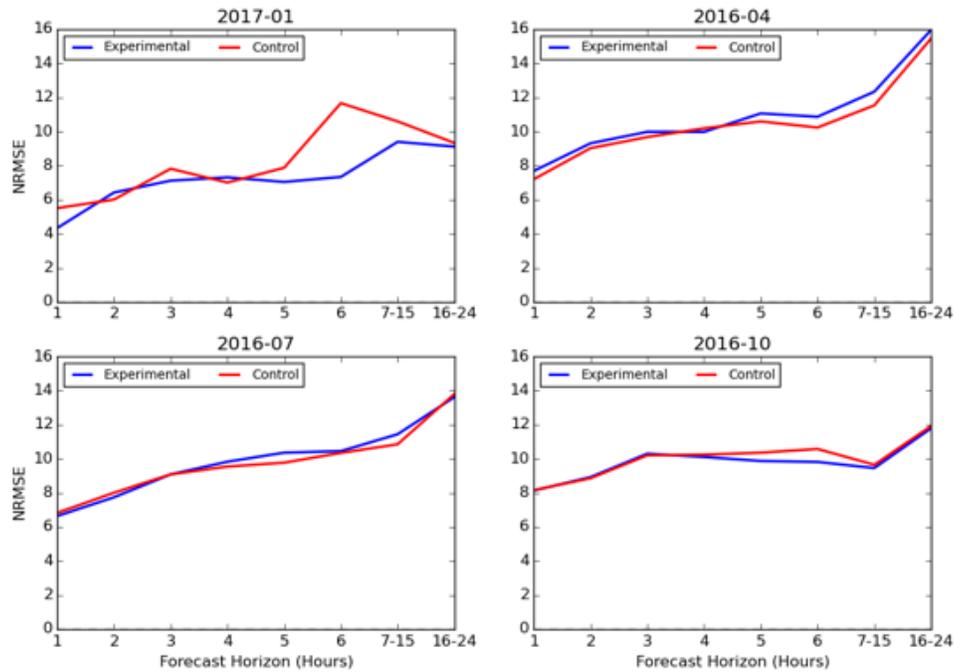


Figure 9. Normalized RMSE (m/s) for wind power forecasts for the BPA fleet aggregate using the experimental (blue) and control (red) 750-m HRRR simulations for the months of January 2017, April 2016, July 2016, and October 2016, grouped by forecast hours 1–6, averaged from 7–15, and averaged 16–24 as indicated on the x-axis

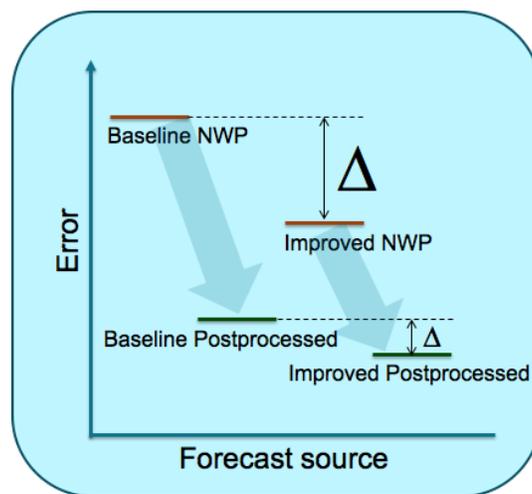


Figure 10. Schematic of error reductions. Improved simulations exhibit lower errors than baseline NWP output. Postprocessing the baseline NWP output reduces the error considerably. Postprocessing improved simulations yields the lowest errors.

4.5 Conference Contributions

Results were presented at conferences (the American Meteorological Society (AMS) annual meetings in Austin and Seattle both had two WFIP 2 sessions). Table 1 lists the latest conference contributions from the 98th annual meeting in Austin.

Table 1. WFIP 2 Related Talks and Poster Presentations at the 98th AMS Annual Meeting in Austin, Texas

Paper Title	Authors
The Wind Forecast Improvement Project 2 (WFIP 2): Overview and Preliminary Model Improvements	Melinda Marquis , NOAA, Boulder, CO; and J. Olson, J. Kenyon, M. D. Toy, Y. Pichugina, A. Choukulkar, T. A. Bonin, R. Banta, L. Bianco, I. V. Djalalova, K. McCaffrey, J. M. Wilczak, K. Lantz, C. N. Long, S. Redfern, J. McCaa, E. P. Gritmit, J. W. Cline, W. Shaw, and J. Sharp
Improvements to Low-Level Wind Forecasts in Complex Terrain from WFIP 2	Jaymes S. Kenyon , NOAA/ESRL and CIRES, Univ. of Colorado, Boulder, CO; and J. B. Olson, M. D. Toy, J. M. Brown, W. M. Angevine, M. Marquis, Y. Pichugina, A. Choukulkar, T. A. Bonin, R. Banta, L. Bianco, I. V. Djalalova, K. McCaffrey, J. M. Wilczak, K. Lantz, C. N. Long, S. Redfern, J. Sharp, J. McCaa, M. T. Stoelinga, E. P. Gritmit, J. W. Cline, W. J. Shaw, J. K. Lundquist, K. A. Lundquist, B. Kosovic, P. A. Jimenez, L. K. Berg, and V. R. Kotamarthi
Identification and Analysis of Forecast Model Large Error Events During WFIP 2	James Wilczak , NOAA/ESRL, Boulder, CO; and K. McCaffrey, I. V. Djalalova, L. Bianco, J. B. Olson, J. Kenyon, M. T. Stoelinga, J. Sharp, M. Pekour, D. Cook, G. Scott, R. Eckman, and S. Otarola-Bustos
Identification and Characterization of Cold Pool Events during WFIP 2	Katherine McCaffrey , CIRES, Boulder, CO; and J. M. Wilczak, L. Bianco, I. V. Djalalova, R. Banta, T. A. Bonin, W. A. Brewer, A. Choukulkar, D. Cook, R. L. Coulter, H. J. S. Fernando, K. Friedrich, L. Leo, J. K. Lundquist, P. Muradyan, and Y. Pichugina
Sensitivity of Turbine Height Wind Speeds to Parameters in Planetary Boundary Layer and Surface Layer Schemes in WRF: Extension to Wintertime Conditions	Larry K. Berg , PNNL, Richland, WA; and Y. Liu, B. Yang, Y. Qian, P. L. Ma, and Z. Hou
Parametric and Initial Condition Sensitivity of Wind Ramp Events	Nicholas H. Smith , Texas Tech Univ., Lubbock, TX; and B. C. Ancell

Paper Title	Authors
Assessment of the WRF Wind Farm Parameterization for Easterly Wake Events during the Second Wind Forecast Improvement Project	Rochelle Worsnop , Univ. of Colorado, Boulder, CO; and J. K. Lundquist, B. Kosovic, P. A. Jimenez, Y. Pichugina, A. Choukulkar, T. A. Bonin, and B. J. McCarty
Boundary Layer Depth Variability Over Complex Terrain as Observed During the WFIP 2	Paytsar Muradyan , Argonne National Laboratory, Argonne, IL; and R. Coulter, R. Kotamarthi, D. Cook, T. J. Martin, H. J. S. Fernando, L. Leo, and S. Otarola
Comparison of Boundary Layer Heights in the Columba River Gorge and Basin from Wind Profiling Radars and Numerical Weather Prediction Models during WFIP 2	Laura Bianco , CIRES, Boulder, CO; and I. V. Djalalova, J. M. Wilczak, K. McCaffrey, J. B. Olson, J. Kenyon, K. Lantz, and C. N. Long
Cloud Fraction and Cloud Base Measurements from Scanning Doppler Lidar during WFIP-2	Timothy A. Bonin , NOAA/CIRES, Boulder, CO; and C. N. Long, K. Lantz, A. Choukulkar, Y. Pichugina, B. J. McCarty, R. M. Banta, W. A. Brewer, and M. Marquis
Surface Solar Radiation Observations at Three Sites across the Columbia River Basin as Part of the Wind Forecasting Improvement Project (WFIP-2)	Kathleen Lantz , NOAA, Boulder, CO; and C. N. Long, A. McComiskey, A. Kumler, M. Marquis, J. B. Olson, J. Kenyon, G. B. Hodges, E. Hall, and J. Wendell
Evaluating Model Skill at Predicting Recurrent Diurnal Summertime Wind Patterns in the Columbia River Basin during WFIP-2	Robert M. Banta , NOAA/ESRL, Boulder, CO; and Y. Pichugina, W. A. Brewer, A. Choukulkar, C. Bonfanti, B. J. McCarty, T. A. Bonin, S. P. Sandberg, J. B. Olson, J. Kenyon, S. Benjamin, K. Lantz, C. N. Long, A. McComiskey, L. Bianco, I. V. Djalalova, J. Wilczak, J. Sharp, D. Cook, R. Eckman, M. T. Stoelinga, J. McCaa, M. Marquis, W. J. Shaw, and J. W. Cline
Advancing NWP Wind Forecasts in Complex Terrain by Lidar Technology	Yelena Pichugina , NOAA/ESRL/Chemical Sciences Division (CSD) and CIRES/Univ. of Colorado, Boulder, CO; and R. Banta, T. A. Bonin, A. W. Brewer, A. Choukulkar, B. J. McCarty, J. W. Cline, H. J. S. Fernando, J. Kenyon, R. Krishnamurthy, M. Marquis, and J. B. Olson
WFIP 2 Experiment: Wind Speed Ramp Events at 80 m above the Ground	Irina V. Djalalova , CIRES, Boulder, CO; and L. Bianco, J. Wilczak, J. B. Olson, J. Kenyon, E. Akish, M. T. Stoelinga, L. K. Berg, D. Cook, R. Coulter, R. Eckman, H. J. S. Fernando, E. P. Gritmit, L. Leo, M. Marquis, P. Muradyan, S. Otarola, M. Pekour, G. Scott, and J. Sharp

Paper Title	Authors
Using Lidar and Nacelle-Mounted Anemometer Measurements to Analyze a Complex Terrain Effect on Wind Flow Pattern	Brandi J. McCarty , CIRES/Univ. of Colorado, Boulder, CO; and Y. Pichugina, A. Choukulkar, T. A. Bonin, A. Brewer, M. Marquis, and J. Sharp
WFIP 2 Validation and Verification Strategy	Caroline Draxl , National Renewable Energy Laboratory, Golden, CO; and R. Banta, L. K. Berg, L. Bianco, T. A. Bonin, A. Choukulkar, A. Clifton, J. W. Cline, E. Colon, I. V. Djalalova, E. P. Gritmit, K. Holub, J. S. Kenyon, B. Kosovic, V. R. Kotamarthi, K. Lantz, C. Long, J. K. Lundquist, J. McCaa, K. McCaffrey, J. F. Newman, J. B. Olson, Y. Pichugina, J. Sharp, W. J. Shaw, N. H. Smith, M. D. Toy, V. Ghate, and R. Worsnop
Overview of WFIP 2 Model Validation Efforts	Melinda Marquis , NOAA, Boulder, CO; and Y. Pichugina, R. Banta, A. Choukulkar, T. A. Bonin, B. J. McCarty, L. Bianco, I. V. Djalalova, K. McCaffrey, J. M. Wilczak, K. Lantz, C. N. Long, J. B. Olson, J. S. Kenyon, M. Toy, K. Holub, R. Pierce, C. Draxl, J. F. Newman, A. Clifton, J. K. Lundquist, R. Worsnop, B. C. Ancell, L. K. Berg, K. A. Lundquist, and J. W. Cline
Can Wakes be Accurately Characterized in Complex Terrain?	Aditya Choukulkar , Univ. of Colorado/NOAA/ESRL/Chemical Sciences Division (CSD), Boulder, DC; and B. J. McCarty, Y. Pichugina, T. A. Bonin, R. M. Banta, S. P. Sandberg, A. Weickmann, and W. A. Brewer
Analysis of Waked Wind Flow in Complex Terrain from Doppler Lidar Measurements	Yelena Pichugina , NOAA/ESRL/Chemical Sciences Division (CSD) and CIRES/Univ. of Colorado, Boulder, CO; and N. Persson, A. Choukulkar, R. M. Banta, T. A. Bonin, W. A. Brewer, B. J. McCarty, S. P. Sandberg, and A. M. Weickmann

5 Summary

The goal of the WFIP 2 V&V effort is to provide tools, methods, and guidance to enable repeatable, metrics-based assessment of the WRF and HRRR models. A comparison of simulations with observations is related to stakeholders in wind energy and targeted at their needs.

Establishing a V&V process with many team members across various labs and entities, as well as multiple funding sources, is challenging at times. This can be alleviated by an interactive V&V effort in which team members shape the process as much as possible. Tools are developed together and in agreement, the methods are discussed at biweekly meetings, and guidance is provided at meetings and as needed. The team also coordinates WFIP 2-related publications and consolidates results.

The key meteorological variables for WFIP 2 are the 80-m or 100-m wind speeds in areas where there are wind power plants. The wind ramp metric is important because it isolates performance evaluation to the periods when wind speed is rapidly changing and can also be compared to the results from WFIP1 where the metric was first devised. RMSE, bias, MAE, % improvement thereof and of the bulk rotor layer wind speeds between 40 m and 200 m, as well as their behavior throughout the day, season, and averaged over areas, are all used within the project.

5.1 Lessons Learned

Initially, the idea that a V&V strategy should be employed within WFIP 2 was doubted by some members; therefore, initial efforts focused on explaining the benefits of a common V&V strategy to the team members. Oberkampf and Roy (2012) compare the adoption of a V&V culture to an adoption of any new culture: Adopting a new culture is difficult and it takes time for people to adjust. After all, we are dealing with humans, their interactions with each other and we leverage their years-long expertise that they don't like being questioned about. Therefore, in WFIP 2 we have moved to an interactive V&V effort where team members shape the process as much as possible. Regular in-person meetings were very helpful to talk over preliminary results, discuss issues and concerns, or discuss further steps.

Initial efforts, such as using a common code base for validation efforts, proved difficult, mostly because a common code base was not available at the start of the project. Therefore, the V&V team organized workshops to test individual validation code on a defined data set and defined problem. This turned out to be extremely crucial. For example, error metrics for a same measurement device at the same location over the same time period were found to have differences of up to 2 meters per second (m/s). Possible reasons for this discrepancy included wrong interpretation of the timestamp of the observations and, thus, a 1-h shift and different averaging techniques in the vertical and horizontal.

The V&V effort has proven useful in coordinating a common messaging at conferences and for publications in general. We wanted to avoid that team members would present preliminary results and possibly contradict each other. We also went a step further and coordinated WFIP 2 related publications and consolidated results.

5.2 Recommendations for the Future

In the future, we recommend that V&V strategies are planned out before the start of a project and that initial high-level V&V planning meetings be held in person. This will allow for a focused audience and beneficial input. Successes in later WFIP 2 V&V meetings can probably be attributed to the setup of in-person meetings at the institution with the most participants (NOAA).

We further recommend that an EMAT or PIRT be developed *before* installing the observations. That way, crucial measurements at important locations can be optimally deployed to help validate model simulations and the use of the observations will be clear during the experiment.

Furthermore, a sub-V&V team should be established at each participating lab or institution. Acceptance and follow through of new strategies and methods is easier in smaller groups. Getting buy-in from a large group can be challenging.

Last, but not least, a V&V process needs a changing mindset. A changing mindset takes time, which is something that should be considered. With this in mind, frustrations can be mitigated and success established.

References

American Institute of Aeronautics and Astronautics' Guide for the Verification and Validation of Computational Fluid Dynamics Simulations. Computational Fluid Dynamics Committee, eISBN: 978-1-56347-285-5. <https://doi.org/10.2514/4.472855>.

Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer. ASME V&V 20-2009.

W. Oberkampf and C. Roy. 2010. *Verification and Validation in Scientific Computing*. Cambridge.

C. Epifanio and R. Rotunno. 2005. The Dynamics of Orographic Wake Formation in Flows with Upstream Blocking. JAS

Bianco, L., I. V. Djalalova, J. M. Wilczak, J. Cline, S. Calvert, E. Konopleva-Akish, C. Finley, and J. Freedman. 2016. "A Wind Energy Ramp Tool and Metric for Measuring the Skill of Numerical Weather Prediction Models." *Wea. Forecasting*, **31**, 1137–1156, <https://doi.org/10.1175/WAF-D-15-0144.1>.

J. Olson, J. Kenyon, I. Djalalova, L. Bianco, D. Turner, Y. Pichugina, A. Choukulkar, M. Toy, J. M. Brown, W. Angevine, E. Akish, J-W. Bao, P. Jimenez, B. Kosovic; Katherine Lundquist; C. Draxl, J. K. Lundquist, J. McCaa, K. McCaffrey, K. Lantz, C. Long, J. Wilczak, R. Banta, M. Marquis, S. Redfern, L. K. Berg, W. Shaw, J. Cline. 2019. "Improving Wind Energy Forecasting through Numerical Weather Prediction Model Development." *Bull. Amer. Meteor. Soc.* Accepted.

Pichugina, Y. L., R. M. Banta, T. Bonin, W. A. Brewer, A. Choukulkar, B. J. McCarty, S. Baidar, C. Draxl, H. J. Fernando, J. Kenyon, R. Krishnamurthy, M. Marquis, J. Olson, J. Sharp, and M. Stoelinga. 2019. "Spatial Variability of Winds and HRRR–NCEP Model Error Statistics at Three Doppler-Lidar Sites in the Wind-Energy Generation Region of the Columbia River Basin." *J. Appl. Meteor. Climatol.*, **58**, 1633–1656, <https://doi.org/10.1175/JAMC-D-18-0244.1>.

Shaw, W. J., L. K. Berg, J. Cline, C. Draxl, I. Djalalova, E. P. Gritmit, J. K. Lundquist, M. Marquis, J. McCaa, J. B. Olson, C. Sivaraman, J. Sharp, and J. M. Wilczak. 2019. "The Second Wind Forecast Improvement Project (WFIP 2): General Overview." *Bull. Amer. Meteor. Soc.*, **0**, <https://doi.org/10.1175/BAMS-D-18-0036.1>.

Wilczak, J. M., M. Stoelinga, L. K. Berg, J. Sharp, C. Draxl, K. McCaffrey, R. M. Banta, L. Bianco, I. Djalalova, J. K. Lundquist, P. Muradyan, A. Choukulkar, L. Leo, T. Bonin, Y. Pichugina, R. Eckman, C. N. Long, K. Lantz, R. P. Worsnop, J. Bickford, N. Bodini, D. Chand, A. Clifton, J. Cline, D. R. Cook, H. J. Fernando, K. Friedrich, R. Krishnamurthy, M. Marquis, J. McCaa, J. B. Olson, S. Otarola-Bustos, G. Scott, W. J. Shaw, S. Wharton, and A. B. White. 2019. "The Second Wind Forecast Improvement Project (WFIP 2): Observational Field Campaign." *Bull. Amer. Meteor. Soc.*, **0**, <https://doi.org/10.1175/BAMS-D-18-0035.1>.