Final Technical Report Cover Page

Federal Agency to which Report is submitted: DOE EERE – Wind & Water Power Program

Recipient: Vaisala Inc.

Award Number: DE-EE0006898

Project Title: Wind Forecasting Improvement Project In Complex Terrain Near the Columbia River Gorge

Project Period: December 23, 2014 – September 30, 2018

Principle Investigator: James McCaa, Senior Scientist, jim.mccaa@vaisala.com, 206-430-4649

Report Submitted by: James McCaa, Senior Scientist, jim.mccaa@vaisala.com, 206-430-4649

Date of Report: Feb 28, 2019

Covering Period: December 23, 2014 – September 30, 2018

Working Partners: University of Colorado (Julie Lundquist, julie.lundquist@colorado.edu)
University of Notre Dame (Joe Fernando, hfernand@nd.edu)
Lockheed Martin Coherent Technologies (Keith Barr, keith.barr@lmco.com)
Texas Technical University (Brian Ancell, brian.ancell@ttu.edu)
National Center for Atmospheric Research (Branko Kosovic, branko@ucar.edu)
Sharply Focused LLC (Justin Sharp, justin@sharply-focused.com)

Cost-Sharing Partners: University of Colorado at Boulder
Boulder, CO 80309-0311

Lockheed Martin Coherent Technologies, Inc. 135 South Taylor Avenue
Louisville, CO 80027-3025

DOE Project Team: DOE HQ Program Manager – Valerie Reed
DOE Field Contract Officer – Laura Merrick
DOE Field Grants Management Specialist – Jane Sanders
DOE Field Project Officer – Brad Ring
DOE/CNJV Project Monitor – Martha Amador

Signature of Submitting Official: 
(electronic signature is acceptable)

Acknowledgment: “The information, data, or work presented herein was funded in part by the Office of Energy Efficiency and Renewable Energy (EERE), U.S. Department of Energy, under the Wind Energy Technologies Office Award Number DE-EE0006898.”

Table of Contents

Executive Summary **4**

1 Introduction **5**

1.1 Role of the Vaisala Team within WFIP2 ..... **7**

1.2 Objectives and Goals of the Project **9**

1.3 Project Management **10**
1.3.1 Steering Committee 10
1.3.2 Teams Structure 10
1.3.3 Communications 10

2 Project Overview **11**

3 Project Results **14**

3.1 Field Campaign **14**
3.1.1 Experimental Design 14
3.1.2 Instrument deployment 17
3.1.3 Event Log 17
3.1.4 Event Catalog and Common Case Study Set 18

3.2 Model Development **20**
3.2.1 3D PBL parameterization 20
3.2.2 Analysis of Wind Farm Wake Parameterization 40

3.3 Observations and Analysis **42**
3.3.1 Observational Processing and Quality Control 42
3.3.2 Analysis of Physics Site Data 42
3.3.3 Uncertainty Quantification 67
3.3.4 Wind Power Forecasts 74

3.4 Decision Support Tools **83**
3.4.1 Methodology 84
3.4.2 Results 90
3.4.3 Prototype Tool 97
3.4.4 Industry Feedback 99
3.4.5 Summary and Discussion 102
3.4.6 Future Work and Recommendations 103

4 Summary **104**

References **106**

List of Acronyms **112**

Executive Summary

In 2014 a multi-institution team led by Vaisala, Inc. was selected by the Department of Energy (DOE) to partner with multiple DOE and National Oceanic and Atmospheric Administration (NOAA) laboratories on a project designed to improve the quality of wind power forecasts in areas of complex terrain. This was the second Wind Forecast Improvement Project (hereafter WFIP2) funded by DOE and it extended from late 2014 through the middle of 2018. It encompassed an 18-month observational field campaign, numerical weather prediction (NWP) model development, extensive analysis of data and NWP output, and the creation of decision support tool algorithms to convey forecast information to end users in the wind industry.

WFIP2 focused on improvements to the representation of near-surface and boundary-layer physics in NOAA's High-Resolution Rapid Refresh (HRRR) model. Improvements to HRRR, which is run operationally over the continental United States, benefit the wind industry in multiple ways. Forecasts from the operational HRRR are used directly by wind power forecast vendors and the operators of wind plants. In addition, because HRRR is built using the widely-used Weather Research and Forecasting (WRF) model, improvements to its parameterizations become available to commercial and research institutions using WRF for a myriad of purposes.

The geographic area studied by WFIP2 was a region of the Columbia River basin located to the east of the Cascade Mountains between Oregon and Washington. Home to over 6 GW of installed capacity for wind energy production, this area also hosts a variety of atmospheric phenomena either unique to or augmented by complex topography. This makes it an attractive test-bed for the analysis of wind forecast in complex terrain, though results found are should be applicable in any area of topographic complexity.

WFIP2 succeeded as a collaborative effort, and while this report focuses on the activities of the team led by Vaisala, the work described here is part of a larger whole. The Vaisala team accomplished a number of specific tasks as described in this report, while also contributing to this larger effort.

The primary accomplishments of the Vaisala team under WFIP2 were:

- Creation of an experimental design for the overall project.
- Logistical arrangements for field study locations.
- Deployment/maintenance/removal of instruments for the field study.
- Analysis of field study data.
- Development of a 3D PBL parameterization for WRF.
- Creation of a data catalog to enhance the value of the field study observations.
- Creation and analysis of historical NWP forecast simulations.
- Generation and validation of wind power forecasts based on NWP model output.
- Creation of decision support algorithms and development of a prototype display.

WFIP2 was conducted in an open and collaborative manner, with data and model improvements shared publicly wherever possible.

1 Introduction

With wind energy now producing more than 6% of electrical generation in the United States, the need for accurate forecasts of wind generation is greater than ever. Operation of the country's electrical infrastructure, as well as efficiency in our energy markets, critically depends on the quality of wind forecasts, and this dependency is expected to grow with increased penetration of wind generation into the mainstream energy industry (Marquis et al. 2011). Forecasts are required on a variety of time horizons, and this project focuses on the time range of 0-15 hours.

Hours-ahead to days-ahead wind power forecasts are typically generated via a combination of data analysis, numerical modeling of the atmosphere, and engineering models to convert simulated weather into anticipated generation. The first two of these are meteorological in nature and highlight two significant challenges to the production of accurate forecasts: estimation of the current state of the atmosphere, and the evolution in time of that state using parameterized equations within a numerical weather prediction (NWP) model.

The gradual buildout of wind energy projects across the United States has increasingly led to development of utility-scale wind installations in areas of complex terrain, and this impacts both of these forecasting challenges. In regions of complex orography, observations are representative of smaller regions of the atmosphere than in flat orography, and also the necessary representation of physics within NWP models becomes more complex as assumptions valid in flat orography no longer hold true. Many atmospheric phenomena such as gravity waves, cold pools, topographic wakes, and gap flows are either unique to or enhanced by complex terrain.

Skillful weather forecasting in complex terrain requires the accurate prediction of phenomena at a range of spatio-temporal scales as well as the ability to correctly diagnose interactions between scales. When represented in NWP models at sufficient resolution, complex terrain provides a forcing that can make the general weather pattern easier to predict; for example, high pressure west of an east-west mountain gap will yield wind that is generally west to east and modified by the terrain. However, complex topography also amplifies errors in the synoptic scale forecast and produces phenomena that are poorly represented in NWP models either due to limited resolution or shortcomings in model physics. Phenomena poorly represented in models include frontal passages with stable mix-out, gap flows, mountain waves, mesoscale topographic wakes, convective outflows, marine pushes, land-sea breezes, slope and drainage flows, and low-level jets. All of these features have two things in common: 1) they are created or enhanced by topography, and 2) their evolution depends on the interaction between the boundary layer and the large-scale flow, a well-known weakness in NWP models, especially during stable conditions.

What would be a relatively simple transition associated with a frontal passage in the Plains States takes on very complex attributes as it interacts with terrain. Surface and upper air features often become decoupled and existing low-level features such as inversions and gap flows may produce results that are counter to those that would occur in less complex terrain. In addition, higher spatial and temporal resolution observations are needed to correctly diagnose phenomena because the area and time over which an observation is representative is reduced, yet often the density of observations in complex terrain is low due to the logistical problems of instrumenting remote locations.

Thus, forecasting wind power generation in a complex terrain environment is very challenging. Small errors in wind speed forecasts are magnified by the cubic relationship between wind speed and power. In simpler terrain, transitions from low to high output and vice versa usually occur smoothly, but in complex terrain these transitions can be punctuated by embedded up and down ramps that are not easily simulated by NWP models. To compensate for this, statistical models are commonly used, especially at

0-6 hour forecast horizons when nearby observations can provide predictive value. Available observations are often too sparse to properly initialize or validate the models. Complex terrain also makes it far more difficult to use statistical methods to correct for model error and thus much more important to correctly diagnose features, especially when one considers North American wind farms are often 200 MW or larger and extend across hundreds of square kilometers. In areas of complex terrain some turbines may be 500 m higher in elevation than others in the same project. Thus, within one project, several key phenomena may influence different parts of the wind farm at the same time and may also interact with each other. For instance, a wake from a large mountain may reduce the output of the southern part of a plant, while channel flow may be enhancing the flow in the northern part, and gravity waves from a nearby ridge might be influencing the flow throughout. All these effects also have non-linear feedbacks upon each other.

Considerable field work and associated analysis and modeling studies have been carried out in the last 15 years to help us understand these various processes and interactions. In the late 1990s, projects such as SABLES-98 (Cuxart et al., 2000) and CASES-99 (Poulos et al., 2002) helped us understand the behavior of the stable and transitioning boundary layer independent of complications of terrain complexity, with heavily instrumented measurement campaigns over relatively flat areas of North America and Europe. These studies provided insights into the structure and evolution of the stable nocturnal boundary layer, including the development of nocturnal inversions, and mechanisms for formation and maintenance of low-level jets. Low-level jets are of particular importance in wind energy, because they are characterized by a vertically localized maximum in wind speed that is typically near the hub height of modern utility-scale wind turbines. Small errors in predicting the height or intensity of low level jets can have profound impacts on the accuracy of wind power forecasts.

Since 2000, several field studies have investigated the rich palette of flow structures created by complex topography, either through thermal forcing (katabatic flows, mountain/valley circulations, etc.) or the dynamical response of air flow to topography (vertically propagating mountain waves, lee waves, gap flows, topographic wakes, etc.) The Vertical Transport and Mixing Experiment (VTMX; Doran et al., 2002) examined topographically influenced flows and boundary layer processes in the Salt Lake Valley in 2000. This study elucidated the structure and evolution of shallow katabatic flows on sloped terrain, and the existence of a mesoscale low-level jet driven by diurnal thermal forcing across multiple basins. Other VTMX studies looked at model performance with different vertical coordinates and boundary layer closures, idealized simulations and theoretical descriptions of katabatic flows, and valley flows. More recent follow-up studies in the same area include an investigation of the Weber Canyon valley flow, examining how valley flows evolve into exit jets at the opening of the valley, and the recent Mountain Terrain Atmospheric Modeling and Observations (MATERHORN; Fernando and Pardyjak 2013) study, which used a variety of instruments in the Dugway Proving Grounds to document complicated microscale topographic flows, and interactions between flows generated by neighboring mountain features. MATERHORN has illustrated microscale features and flow interactions that cannot be explicitly represented in current models with parameterized boundary layers, punctuating the challenge of improving such models.

Closer to the WFIP2 area, the Columbia Basin Wind Energy Study (CBWES; Berg et al., 2012) was carried out from 2010-2012, on Vansycle Ridge, a wind energy producing region approximately 30 km southeast of the confluence of the Snake and Columbia Rivers. This study measured the vertical profile of wind across the boundary layer over the course of several months, and analyzed the occurrence of hub-height wind ramps and their predictability with the WRF mesoscale model. The study found that in stable conditions, up-ramps were associated with rapid accelerations of nocturnal low-level jets as the boundary layer decoupled from the surface. These wind ramps generally formed under weak synoptic-scale forcing, but may have been influenced by terrain-induced pressure gradients. In unstable conditions, up-ramps were categorized as those influenced by a strong upper-level trough, those with a low-level jet driven by synoptic forcing, those associated with frontal passages, and those influenced by small-scale variations that are probably thermally or terrain induced.

Some studies have examined the role of dynamically driven topographic flows such as mountain/lee waves, gap flows, and topographic wakes on near-surface winds within complex terrain, especially in terms of interactions with stable and transitioning boundary layers. The Terrain-Induced Rotor Experiment (T-REX; Grubišić et al., 2008) examined mountain wave dynamics, but focused on more extreme downslope windstorms and rotors in the lee of the Sierra Nevada mountain range barrier. Observational and modeling studies of the gap-flow dynamics within the Columbia River Gorge (Sharp and Mass 2002, 2004), overlapping the proposed study area, have elucidated the relationship between the within-gap flow and the acceleration at the gap exit region.

As discussed in Shaw et al. (2019), the DOE has been actively supporting the advancement of wind power forecasting for a number of years. As far back as the 1970's, DOE's Atmospheric Studies in Complex Terrain (ASCOT) program funded a series of field studies that analyzed boundary layer flow over complex topography (Orgill and Schreck 1985, Clements et al. 1989, Coulter and Martin 1996). More recently, a DOE wind industry workshop held in 2008 (Schreck et al., 2008) identified the need for NWP forecast improvement, and was followed by the first Wind Forecast Improvement Project (hereafter WFIP 1) in 2011-12 (Wilczak et al., 2015). WFIP 1 addressed the forecast initialization problem by examining the effects of assimilating improved observations on forecast accuracy. A second DOE workshop in 2012 focused on complex atmospheric flows and identified four broad scales of processes that influence wind generation. These were described as volumes of the atmosphere containing turbine scale, wind power plant scale, mesoscale and global processes. Mesoscale processes, influenced by the global and plant scales, drive the variability of wind generation on the hours-ahead to days-ahead time horizon.

The importance of mesoscale processes combined with the increase of wind plants located in complex terrain led the DOE in 2014 to initiate a second Wind Forecast Improvement Project (hereafter WFIP2), this time focused on observing and improving the modeling of flow in complex terrain by better representing boundary layer physics and related processes in a mesoscale model. WFIP2 was conducted as public-private partnership under the DOE's Atmosphere to Electron (A2e) Initiative. Direct involvement of the wind industry both facilitated access to observational data from wind plants and helped guide the project toward solutions relevant to wind plant operations.

WFIP2 extended through four calendar years and included three distinct but closely related components: an 18-month multi-scale field study, model development work, and the development of support tools to support the industry in wind power forecasting. Four associated journal papers describe the activities of the broad WFIP2 team: Shaw et al. 2019 introduces the project, Wilczak et al. 2019 presents an overview of the observational campaign, Olson et al. 2019 presents the model development effort, and Grimit et al. 2019 presents the development of decision support tools. These papers provide valuable context for the work conducted by the Vaisala team that is described in this report.

1.1 Role of the Vaisala Team within WFIP2

The complete WFIP2 project team included DOE and NOAA labs as well as a team led by Vaisala that further included multiple universities, wind industry participants, and the National Center for Atmospheric Research (NCAR). We refer to the Vaisala-led component of the overall team as the 'Vaisala Team' throughout this report and aim to highlight the activities and contributions of this team within the context of the overall WFIP2 effort. The institutional members of the Vaisala Team, along with primary contacts, are listed in Table 1.

Table 1. Members of the Vaisala WFIP2 Team

Vaisala Team Members	Primary Contacts
Lockheed Martin, Inc.	Keith Barr, keith.barr@lmco.com
National Center for Atmospheric Research (NCAR)	Branko Kosovic, branko@ucar.edu
Sharply Focused, LLC	Justin Sharp, justin@sharply-focused.com
Texas Technical University (TTU)	Brian Ancell, brian.ancell@ttu.edu
University of Colorado (CU)	Julie Lundquist, julie.lundquist@colorado.edu
University of Notre Dame (UND)	Joe Fernando, hfernand@nd.edu
Vaisala, Inc.	Jim McCaa, jim.mccaa@vaisala.com

The Vaisala Team had a unique role within the overall project, having proposed the study area in its application for a funding opportunity award, and representing the only industry participants in the project. Our rationale for proposing the Columbia Basin was twofold: it is both meteorologically interesting and important to the wind industry. The region separates Washington and Oregon and is an exceptional natural observatory for studying meteorological phenomena associated with complex terrain. It is most well-known for the near sea-level gap that takes the Columbia River through the Cascade Range, a barrier that averages about 5000 to 6000 feet high. However, the terrain complexity goes well beyond the steep walls surrounding the Columbia River. The Cascade Range itself contains several elevated gaps, and the volcanoes of Mount Hood (3,429 m; 11,250 ft) and Mount Adams (3,743 m; 12280 ft) tower above the near sea-level valleys to the east and west. The canyon carved by the Columbia River continues eastward for over a hundred km east of the Cascade Crest, and numerous feeder channels are formed by its tributaries. The vast Columbia Basin to the east is formed by mountains on all four sides. The properties of the two air-sheds are often radically different, yielding large gradients in temperature and pressure in the lower atmosphere. Air may be guided downslope and channeled, piled up in natural reservoirs, or forced up over more stable air below depending on atmospheric conditions. The results are cold pools, gap flows, mountain waves, mountain wakes, downslope flows, mountain-valley circulations and every manner of other terrain driven circulation at every scale and varying depending on the atmospheric conditions.

At the same time as having more than its fair share of important atmospheric phenomena, the region is also home to over 6 GW of wind energy capacity. Not only does the region experience exceptional wind resource, but also benefits from substantial federally-funded electric transmission infrastructure built to support the Columbia River's significant hydroelectric resource. Numerous wind plant operators (listed in Table 2) chose to join the WFIP2 effort by providing valuable data to the project.

Table 2. WFIP2 Data Partners

Wind Industry Partners
Avangrid
Eurus Energy
NextEra Energy
Portland General Electric
Siemens Heavy Industries
Southern California Edison Company
White Creek Wind

While credit for the successes of WFIP2 must be distributed among all of its members, the Vaisala Team played a key role in maintaining an industry focus for the larger team's activities.

1.2 Objectives and Goals of the Project

The overall mission and first goal of the WFIP2 project was to quantitatively improve forecast skill through better understanding and representation of the physics in the foundational models comprising an NWP-based forecasting system. Working in complex terrain meant that the project first needed to characterize physical phenomena, processes, and the atmospheric properties that occur in regions of complex topography. Particular focus was to be spent on phenomena likely to impact wind speeds and direction at the hub heights of wind turbine generators.

A second goal of the project was to develop new or improved mesoscale model schemes or atmospheric modeling to better represent the identified physical processes. Increasing the accuracy of time-averaged wind and wind variability in the 0 to 15 hour forecasts should have positive implications for day-ahead forecasts made with these foundational weather models.

A third goal was to develop prototype decision support tools that could convey probabilistic forecast information, uncertainty quantification and forecast reliability for system operations.

As members of the larger WFIP2 effort, the Vaisala Team had specific objectives within the overall project. These included:

1. Design and implementation of an observational field campaign in the Columbia Gorge area to detect and analyze phenomena that cause wind variability in complex terrain.
2. Identification of the diverse physical phenomena, processes, and atmospheric physics in the study area.
3. Improvement of surface layer and planetary boundary layer (PBL) parameterizations within the Weather Research and Forecasting (WRF) model.
4. Development of decision support tools to provide enhanced short term forecast context and uncertainty quantification, in direct collaboration with wind facility and electric system operators.
5. Dissemination of verified decision support algorithms, contribution of open source code to the WRF-ARW model, and distribution of project data.

These objectives were of course pursued in close collaboration with the full WFIP2 team, but are focal points for our analysis of our contributions to the project.

1.3 Project Management

The overall WFIP2 team had over a dozen participating institutions and over one hundred active participants and required a large enough project management effort that we mention here four effective tools that helped keep things on track.

1.3.1 Steering Committee

The WFIP2 steering committee was the main decision-making body for the project, and was comprised of five members: one each representing the DOE laboratories, the NOAA laboratories, and the Vaisala Team, and in addition a representative from the DOE contract office and from DOE headquarters. Steering committee meetings were held bi-weekly through the project, with activities that ranged from decisions on spending and technical matters to cross-institution presentations of scientific and technical research developments.

1.3.2 Teams Structure

Under the steering committee, the WFIP2 activities were undertaken by a variety of topical teams with multi-institutional representation on each. These included teams for each of the following:

- Experimental Design
- Instrument Deployment, Maintenance, and Monitoring
- Data
- Verification and Validation (V&V)
- Uncertainty Quantification (UQ)
- Model Development
- Decision Support Tools (DST)

Each team had a formal charter with defined responsibilities and reported to the steering committee.

1.3.3 Communications

The project made extensive use of Google Drive and Google Groups to facilitate sharing of information, data, and presentation materials. We sought to provide a way from people across multiple organizations to collaborate effectively and securely, with the following goals:

- Ensure emails reach the right people by emailing teams rather than individuals.
- Provide filtering and digest options to manage email volume.
- Promote online collaboration in the drafting of docs, spreadsheets and slides with concurrent access and updates, tracking of changes, and version control.
- Create an online repository for project documents accessible by anyone on the team from anywhere with an internet connection.
- Use groups to enable dynamic team membership that is easily updated.

Most people and organizations attached to the project were able to adopt without significant effort, though some experienced limited access, and group communication was a paradigm shift for some.

1.3.3.1 DOE's Data Acquisition and Portal (DAP)

The entire WFIP2 project shared data effectively between each other and the public via the Data Archive and Portal, a facility that is part of DOE's Atmosphere to Electrons (A2E) initiative. At the end of the project, WFIP2 had 280 unique datasets on the DAP, making up over 200 TB of data. See https://a2e.energy.gov/about/dap for more information.

2 Project Overview

The WFIP2 team approached the improvement of forecasting in complex terrain through an integrated observational and modeling approach. This was driven by the desire to deliver applicable results that could be rapidly deployed to operational weather forecasting while being informed by a data-driven research campaign. With this in mind, two main scientific goals were identified:

1. Improvement of the physical understanding of atmospheric processes that directly impact wind power forecasts in areas of complex terrain.
2. Incorporation of the new understanding into a foundational weather forecasting model in order to improve wind power forecasts.

These goals served to guide the project's field campaign and model development effort. The observational campaign needed to be multi-scale to be able to capture the wide variety of scales associated with the physical processes the control wind speeds across wind plants. Based on the resulting observations, representation of physical processes in model parameterizations were modified. The results of the implemented changes were evaluated through assessment of the improvements in forecast skill, and decision support algorithms were developed to facilitate the operational use of improved forecasts.

In order to capture the large and small scale features that determine the horizontal and vertical profiles of wind speed and turbulence, the domain for the WFIP2 observational campaign stretched from the Pacific Ocean to eastern Washington. As described below, a series of observational 'nests' measured finer scales all the way down to that of a single mesoscale model grid cell. Suites of instruments contributed by DOE and NOAA laboratories as well as the Vaisala Team were deployed to provide complementary data streams, and these were used throughout the project. A project web site set up by NOAA was integral to facilitate real-time comparison of observations and model output.

An early decision was to target model improvements for NOAA's High Resolution Rapid Refresh (HRRR) model, which led to a variety of opportunities and constraints. NCEP's operational HRRR output is widely used for wind power forecasting, both as a direct source of hub height wind fields, and as an input to custom NWP simulations run by industry forecast providers. In addition, HRRR is an implementation of the widely used Weather Research and Forecasting model Advanced Research WRF (WRF-ARW). As such, improvements to HRRR are available to the broad WRF research community.

WFIP2 concentrated on forecasts with lead times of 0-15 hours, with the expectation that physics-based improvements in this time frame should carry over to day ahead forecasts. The operational HRRR is run hourly over the continental United States at a horizontal resolution of 3km. In addition, during the period of WFIP2 NOAA's ESRL laboratory ran a special 750 m nest over the WFIP2 study area.

Targeting a model used for many additional purposes beyond wind power forecasting meant that improvements for the sake of wind power forecasts could 'do no harm' to other model fields and were broadly applicable in all weather conditions. This influenced the requirement that the observational campaign collect data through a full annual cycle.

Observational data were used throughout the WFIP2 project to both inform model development and validate model output. Near the end of the project, a full year of control and experimental historical forecasts were created to facilitate a formal comparison.

As a partner to the DOE and NOAA laboratories, the Vaisala Team assumed responsibility for a specific set of tasks, which are summarized in Table 3. Almost all of these tasks were completed in close coordination with DOE and NOAA, making it a challenge in some areas to separate out Vaisala Team accomplishments from those of the larger project. We consider this a positive outcome.

The following sections of this report describe in more detail the results of the Vaisala Team effort in the field campaign, model development, model analysis, and decision support tool efforts.

Table 3. Vaisala Team Tasks

Task	Summary	Notes
1	Plan for instrumentation layout and deployment schedule for the field study	Planning for the field study occupied the early months of the project and resulted in an experimental design document and a deployment plan. Both of these are available on the DAP.
2	Logistical arrangements for field study sites	Responsibilities for the logistical arrangements for field study sites were shared with the broader WFIP2 team. Each site with multiple instruments had a coordinator to facilitate organized communication between the various institutions deploying instruments. Vaisala was responsible for the Physics Site.
3	Deployment of instruments	The Vaisala team was responsible for the deployment of its instruments. Vaisala Team instruments are discussed in Section 3.1.2.
4	Management of field study	Members of the Vaisala Team played an active role throughout the field study in monitoring the flow of data from the field, analyzing current weather, and coordinating maintenance of equipment. See Section 3.1.3 for a discussion of the Event Log.
5	Model development	NCAR was the primary participant from the Vaisala Team in model development. Section 3.2.1 discusses their development of a 3D PBL parameterization.
6	Model validation and improvement	Activities under this task included CU's analysis of a wind plant wake model (Section 3.2.2) and TTU's work on uncertainty quantification (Section 3.3.3).
7	Catalog of phenomena observed during field study	An annotated list of important meteorological events observed during the field campaign and a set of common case study dates were created. See Section 3.1.4
8	Selection of improved model physics suite	This was an effort led by NOAA with participation from the entire WFIP2 team. Results are discussed in Olson et al. (2019).
9	Evaluation of model improvements within HRRR	In support of this activity, Vaisala worked with NOAA and the Argonne Leadership Computing Facility (ALCF) to complete a full year of retrospective control and experimental forecast simulations on the full 3km HRRR domain as well as the 750m WFIP2 nest. All model output was delivered to the DAP.
10	Wind power forecast generation	Vaisala processed the output from the retrospective simulations into wind power forecasts, as discussed in Section 3.3.4.
11	Forecast validation	Validation of the wind power forecasts is discussed in Section 3.3.4.2.
12	Decision support tool development	The Vaisala Team worked with WFIP2 and industry partners to develop prototype decision support tools to present actionable information from power forecasts, as discussed in Section 3.4.
13	Final Vaisala team project report	Vaisala was responsible for the creation of this report.

3 Project Results

3.1 Field Campaign

The overarching goal of the observational campaign was to provide insights into the relevant meteorological phenomena in order to improve wind power forecasting. Because the targeted phenomena involve the interaction of the large-scale flow with the boundary layer and the surface itself, it was necessary to capture salient features ranging in size from the meso-beta (20-200 km scales) through the meso-gamma (2-20 km) to the microscale (< 1 km). Also, since some of the targeted phenomena are strongly seasonal, a full annual cycle was measured to capture both typical and unusual phenomena that affect wind forecasts throughout the year.

The campaign was tightly integrated across the full WFIP2 team, with a single shared experimental design, and many observational sites featured colocated instruments from multiple institutions. Instruments were in the field for 18 months, from Oct 2015 through March 2017. The field campaign is discussed in detail in Wilczak et al. (2019) and here we will focus on the contributions of the Vaisala team to the larger effort.

3.1.1 Experimental Design

The field study design strategy is best understood by zooming in from the broadest scale to the smallest. At the regional scale, a concentration of wind farms exists in the eastern exit region of the Columbia Gorge, but much of what drives winds east of the Cascades originates to the west, so it was important to identify and quantify approaching frontal systems and upper-level troughs; collect upstream tropospheric vertical profiles of wind and stability that affect the gravity wave response to flow across the Cascade barrier; determine marine air mass conditions prior to and during onshore push events; and monitor convective systems that developed over or west of the Cascade crest and then propagated eastward.

Within the large energy-producing area to the east of the Cascades, the large-scale flow encounters an area of considerable terrain complexity, highlighted by the broad east-west valley of the Columbia River itself and a series of north-south canyons and valleys through which flow tributaries of various sizes. The wind energy plants are generally located on the elevated terrain between these valleys (see Figure 1), in a semi-arid agricultural zone. In this area, a higher concentration of instruments were deployed in a ‘supersite’ in order to observe the interaction of the large-scale flow with the boundary layer and surface.

Figure 1. Typical terrain of wind energy plants in the WFIP2 region. Photo courtesy Justin Sharp.

The general layout of the instrument deployment is shown in Figure 2. Because the observing strategy needed to capture features across such a broad range of atmospheric scales, a set of cascading nests of instruments were deployed:

1. One set of instruments was be placed at meso-alpha scale (200-2000 km) distances from the areas of concentrated wind energy production. These provided insight into large-scale features and helped determine whether they were accurately represented in the HRRR model output.
2. An intermediate-scale set of instruments captured meteorological processes affecting boundary-layer winds over a variety of terrain types, including locations both east and west of the Cascade Mountains.
3. A fine-scale set of instruments was focused on a high concentration of wind energy plants in the central part of the domain.
4. Finally, a 'physics site' (see Figure 3 below) deployed a variety of near-surface instruments in an effort to characterize interactions of the large-scale flow with terrain at scales normally within in single grid cell of an NWP simulation.

Visible in Figure 2 are the north-south feature of the Cascade Mountains through which the Columbia flows, along with the locations of WFIP2 instrument clusters.

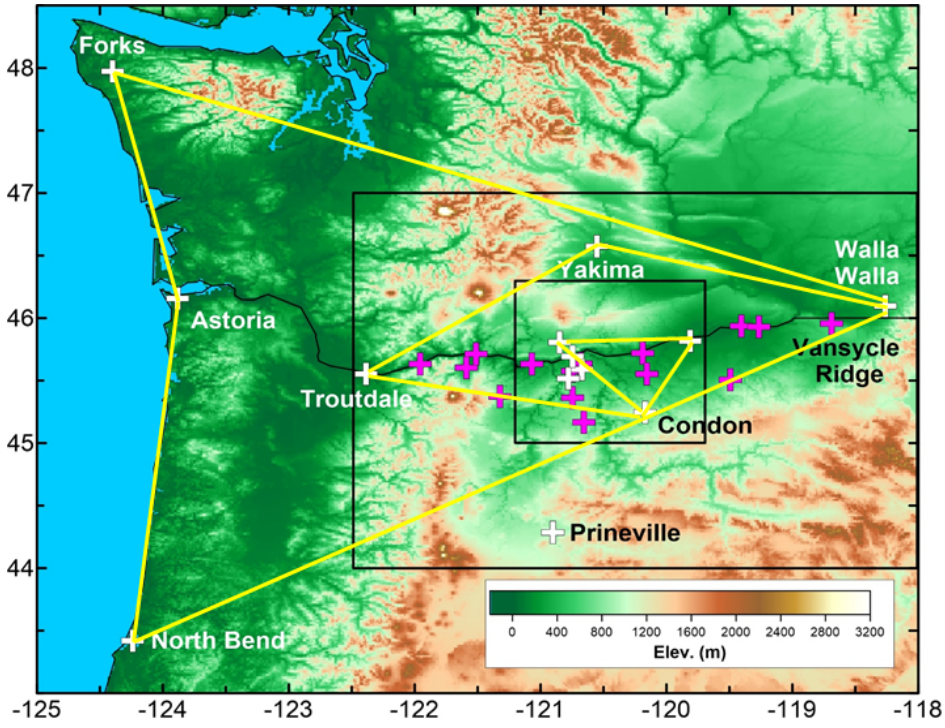

Figure 2. Locations of WFIP2 instruments. Yellow polygons depict the cascade of nesting scales. White crosses indicate instruments, and magenta crosses indicate multi-instrument locations.

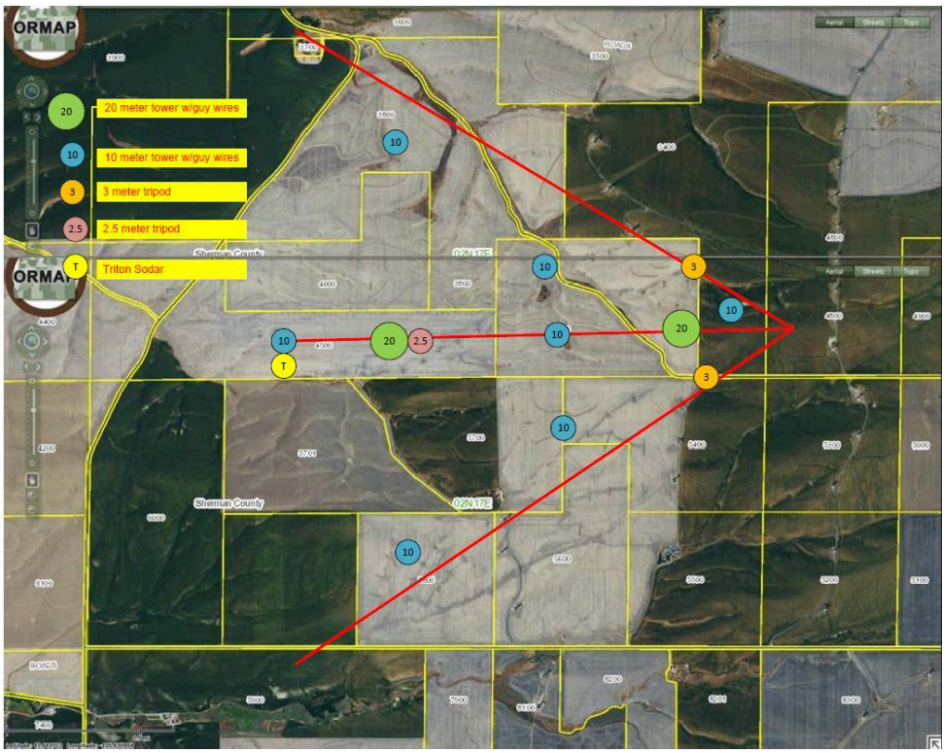

Figure 3. Layout of the physics site. Yellow lines represent distinct land ownership.

3.1.2 Instrument deployment

Vaisala Team equipment was deployed for WFIP2 as part of the overall experimental strategy, and as a result was widely distributed across the region, as shown in Figure 4.

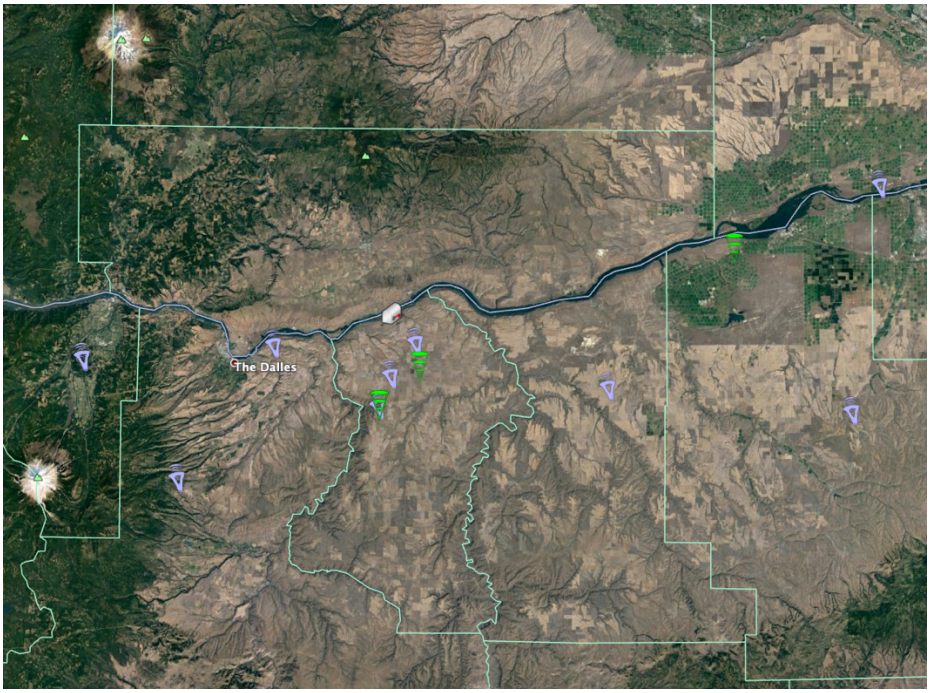

Figure 4. Locations of Vaisala Team SODARs (lavender icons), LIDARs (green icons), and radiometer (mailbox icon).

The full set of Vaisala Team equipment deployed to the field is listed in Table 4, which shows the type of instruments, locations, and dates of deployment and removal from the program. It should be noted that a total of eight Tritons were deployed – after repeated vandalism, one of the eight was relocated to a ninth site. Another had to be replaced after being consumed in a wildfire.

The Vaisala Team played a large role in the logistics for instrument deployment for the larger WFIP2 team, and in particular was responsible for the arrangement of land leases for the Physics Site, as seen in Figure 3.

All data collected during the WFIP2 campaign was transferred to the DOE's Data Archive and Portal (DAP) for use by WFIP2 and other researchers. See https://a2e.energy.gov/projects/wfip2 for more information.

3.1.3 Event Log

A critical task in establishing the usefulness of the WFIP2 observational dataset was documenting the meteorological events captured within it. Both a daily log of events, as well as curated collection of 'important' events were assembled and delivered to DOE as part of the project.

Each week for the duration of the field program, WFIP2 scientists along with other industry participants engaged in a conference call to review the previous 7 days of weather in the region. In addition to the field campaign observations, model forecasts from the real-time 13-km RAP, 3-km HRRR and 750-m WFIP2 HRRR-nest models were discussed. Comparisons were made to time series of aggregate wind power generation within the regional grid operator's (Bonneville Power Administration - BPA) balancing area to establish phenomena that were leading to various levels of forecast skill. Additional data products

that proved useful were global model forecasts, the University of Washington regional forecast model, satellite imagery, local NWS soundings, and ASOS/AWOS surface observations.

For each day of the 18-month field campaign, a brief synopsis of the weather was written, assessing the significance of the key phenomena that impacted wind power generation, along with success of the forecast for that day. Active phenomena including cross-barrier westerly flow, cold pools, mountain waves, topographic wakes, convective outflows, and easterly flow were identified. The cross-barrier flow type was further subdivided into cases forced by low-level mesoscale effects, and those forced by synoptic features, with the latter further subdivided into evolving, mature, and decaying. Cold pools were characterized as deepening, steady state or decaying due to warm advection and mixdown from aloft, cold advection aloft or insolation eroding them from below.

The event log provides a comprehensive list of weather events that occurred during the field campaign along with an assessment of whether those events were important for wind power forecasting. After the completion of the event log, and second, more selective listing of events of particular interest to the projects was compiled as a catalog to assist future researchers. These documents are available on DOE's Data Archive and Portal (DAP).

3.1.4 Event Catalog and Common Case Study Set

At the end of the field study, the Vaisala Team organized the analysis of meteorological events that were recorded in the Event Log through the full observational period to identify key atmospheric structures and processes of interest to the project and likely to expose atmospheric physics improvements in the model PBL schemes, from both a phenomenological and statistical perspective. Each season of the year was evaluated, and both canonical and marginal events were identified. The goal was to form a set of annotated dates and events that could be used to improve understanding of phenomena, increase the ability to predict them, and improve the ability to model those events through improved modeling or PBL schemes. The dates and times identified became the Common Case Study Set which was used to highlight representative forecast skill during important times, and the annotated list of them is the Event Catalog, which was delivered to the DAP and will hopefully be of assistance to future researches in effectively utilization WFIP2 data.

Table 4. Instruments deployed by the Vaisala Team

Instrument Class	Location	Instrument	Owner	Install Date	End Date
LIDAR	Gordons Ridge, OR	Leosphere v2 profiling lidar	CU	11/16/2015	3/13/2017
LIDAR	Troutdale, OR	Leosphere v1 profiling lidar	CU	11/18/2015	1/27/2017
LIDAR	Wasco, OR	Leosphere v1 profiling lidar	CU	2/23/2016	1/27/2017
Microwave Radiometers	Condon, OR	Radiometrics MWR-3000A	CU	11/17/2015	1/2/2017
LIDAR	Gordons Ridge, OR	WindTracer scanning lidar	LMCO	11/19/2015	3/12/2017
Ceilometer	Wasco, OR	Vaisala CL31	UND	12/2/2015	1/25/2017
LIDAR	Boardman	Halo Photonics Stream Line Scanning Doppler LiDAR	UND	12/4/2015	1/25/2017
Microwave Radiometers	Rufus, OR	Radiometrics MWR-3000A	UND	11/19/2015	1/25/2017
SODAR	Rufus, OR	scintec mfas	UND	12/1/2015	1/27/2017
Sonics 10m tower	PS-01	Sonic 10m tower (sonic, T,RH @3m,10m, Licor @3m)	UND	2/15/2016	8/1/2017
Sonics 10m tower	PS-11	Sonic 10m tower (sonic, T,RH @3m,10m)	UND	2/15/2016	8/1/2017
Sonics 20m tower	PS-02	Sonic 20m tower (sonic @3m,10,17m; T,RH @3m,17m)	UND	2/15/2016	1/25/2017
Sonics 20m tower	PS-06	Sonic 20m tower (sonic @3m,10,21m; T,RH @3m,21m)	UND	2/15/2016	1/25/2017
SODAR	15 Mile Road	Triton-AON7	Vaisala	10/1/2015	7/31/2017
SODAR	Gilhouley Road	Triton-AON8	Vaisala	12/7/2015	4/9/2016
SODAR	Gordon Ridge	Triton-AON6	Vaisala	10/1/2015	7/31/2017
SODAR	Old Tree Road	Triton-AON2	Vaisala	10/1/2015	7/31/2017
SODAR	Plymouth	Triton-AON3	Vaisala	10/1/2015	7/31/2017
SODAR	PS01/Scott Canyon Rd	Triton-AON4	Vaisala	12/5/2015	7/31/2017
SODAR	Sand Hollow Road	Triton-AON1	Vaisala	10/1/2015	7/31/2017
SODAR	Shell Rock Road	Triton-AON9	Vaisala	11/19/2016	7/31/2017
SODAR	Van Gilder Road	Triton-AON5	Vaisala	10/1/2015	7/31/2017

3.2 Model Development

Atmospheric flows in complex terrain include many phenomena in which the assumption of horizontal homogeneity implicit in many model parameterizations is not valid. These include slope flows, gravity waves, topographic wakes, thermally-driven valley circulations, gap flows, and cold pools. Even synoptically-forced flows are influenced by local terrain and variations in land cover which lead to heterogeneous surface fluxes of heat and moisture. Modeling such non-stationary flows requires consideration of the turbulent mixing of state variables through the horizontal components of turbulence, which can be as large as the vertical in some cases. In addition, conventional assumptions of horizontal surface homogeneity, stationarity and a constant flux layer near the surface are no longer valid. Even if model resolution is increased to better represent the surface variability, the numerics of mesoscale models can become less accurate as model resolutions approach the dominant turbulence length scales (the “terra incognita” of Wyngaard 2004). The accuracy of prediction of hub height wind speed is significantly affected by the accuracy in representing the effect of boundary layer turbulence on the mean velocity field.

3.2.1 3D PBL parameterization

As a key contribution to WFIP2, NCAR developed and implemented a new PBL parameterization in the WRF-ARW that does not rely on assumptions that limit its accuracy in simulations of flows in complex terrain.

The WRF model is a community-developed, public-domain, mesoscale NWP system designed to serve both operational forecasting and atmospheric research needs for multiple applications (Skamarock 2004; Skamarock et al. 2005; Skamarock and Klemp 2008). The High- Resolution Rapid Refresh (HRRR) NWP framework under development by National Atmospheric and Oceanic Administration (NOAA) uses WRF-ARW to produce hourly forecasts over the Continental United States (CONUS) at three-kilometer resolution (Benjamin et al. 2015). The HRRR is an important component of any short-term wind power forecasting system. NCAR is among the main contributors to the development of WRF-ARW. WRF-ARW includes multiple options for parameterization of physical processes in the atmosphere including parameterization of the effects of unresolved complex terrain.

Improving NWP models in complex terrain requires improved understanding of flows and processes in complex terrain. High-resolution, high-quality flow observations are essential in to achieve better understanding of flows in complex terrain. However, in various terrain induced flow phenomena represent a challenge not only for wind forecasting, but also for observations and analysis. These phenomena include: mountain waves, topographic wakes, gap flows, clod pools, drainage flows, etc. Until recently there were only a few filed studies that comprehensively addressed complex terrain flow phenomena. The WFIP2 project was designed with a goal to provide information needed to improve wind forecasting. To that end Vaisala, Inc. team, together with collaborators, designed a year-long observational field study in the Columbia River Gorge. The Columbia River Gorge was selected for the filed study because it hosts one of the world’s largest concentrations of wind turbines, nestled in an area of extremely complex terrain. A year-long observational field study provided a wealth of data for better characterization of the phenomena that undermine wind forecasts.

To tackle challenges presented by developing better understanding of flows in complex terrain that will lead to better prediction needed are:

- approaches, methods, and technologies for observing flows in complex terrain,
- analysis tools for characterizing complex flows,
- metrics for quantitative characterization of complex flows,
- parameterizations of physical processes in NWP models,
- ways of assessing model performance, and

- standards for acceptable error and uncertainty levels.

Addressing each of these challenges requires a careful consideration and an extensive study. Previous studies that addressed these challenges include Vertical Transport and Mixing (VTMX, Doran et al. 2002), MAP (Rotach and Zardi 2007), T-REX (Grubišić et al. 2008), METCRAX (Whiteman et al. 2008), and MATERHORN (Fernando et al. 2015). The WFIP2 project was designed to address most of them. In this project NCAR team primarily focused on development of a new turbulence parameterization for high-resolution, sub-kilometer grid cell size, mesoscale simulations. The WFIP2 filed study data were used to assess the performance of the new parameterization.

Wind forecast in an atmospheric boundary layer and therefore at a hub-height of a utility scale wind turbine is significantly affected by shear and critically depends on the accuracy of the parameterization of turbulent stresses and fluxes. In NWP models evolution of a flow filed including: velocity, potential temperature, moisture and other constituents is represented using Reynolds Averaged Navier-Stokes (RANS) equations (see Appendix). The RANS equations are usually truncated at some higher moments of field variables and therefore do not represent a closed system of equations. In order to close the system of equations it is necessary to provide parameterizations of higher order moments. In NWP models usually second moments, i.e. turbulent stresses and fluxes, must be parameterized. Until recently NWP models utilized computational grids with grid cell sizes of ten or more kilometers. Recently, continuous development of high-performance computing (HPC) platforms enabled increased resolution of NWP models and grid cell sizes of one kilometer or less. For wind energy applications high resolution mesoscale simulations are essential to capture details of the flow within the first hundred meters above the surface where wind turbines operate. High-resolution simulations are of particular importance in complex terrain where the topography and terrain heterogeneity affect flow evolution and determine flow features at smallest resolved scales. However, in order to improve NWP models for forecasts in complex terrain there are a number of challenges that need to be addressed.

Turbulence parameterizations in mesoscale models are based on volume averaged RANS equations. As the grid cell volume decreases below characteristic atmospheric boundary layer length (ABL) scale, e.g. boundary layer height, the assumptions commonly used to develop these parameterizations are violated. The range of scales of motion between the characteristic ABL scale (approximately one kilometer) and the inertial range of turbulence (less than 100 m in ABLs), where large eddy scale (LES) parameterizations can be used, is often labeled a “grey zone,” or according to Wyngaard (2004) “terra incognita.” Recently a number of modifications to existing parameterizations were proposed in order to develop a “scale aware” parameterization that could seamlessly be used across a range of scales, down into the “terra incognita.” While Wyngaard (2004) proposed to replace a scalar eddy diffusivity with a diffusivity tensor, others focused on modifying length scales used in one-dimensional planetary boundary layer (1D PBL) parameterizations (Efstathiou and Beare 2015; Efstathiou et al. 2016; Shin and Dudhia 2016).

Beare (2014) as well as Honnert and Masson (2014), on the other hand, attempted to estimate the length scale below which a 3D parameterization would need to be used. Honnert and Masson (2014) determined that the exact value varies significantly depending on whether convection is free or forced. Boutle et al. (2014) adopted a blending approach to the grey zone parameterization by combining a 1D PBL parameterization with a subgrid model commonly used in LES.

While a flow over a flat, uniform terrain can be considered horizontally homogeneous, a flow over complex terrain induces circulations resulting in horizontal inhomogeneities. However, at present, most NWP models use 1D PBL parameterizations based on the assumption of horizontal homogeneity. Such models do not account for a range of processes that control production, redistribution, transport, and dissipation of turbulent kinetic energy (TKE) and that in addition to vertical shear and potential temperature gradients also depend on horizontal shear and horizontal potential temperature gradients. In addition, such parameterizations commonly assume that turbulence is in a local equilibrium, i.e. that local production and dissipation of turbulence are in balance. The effects of horizontal gradients of wind and potential temperature are more pronounced at smaller scales and therefore, as the resolution increases,

neglecting these effects result in an inaccurate representation of turbulent stresses and fluxes and consequently errors in a prediction of mean fields. As the NWP model simulation resolution increases it is therefore essential to account for the effects of horizontal shear and temperature gradients on the production and evolution of turbulence including turbulent stresses and turbulent fluxes. We have therefore developed a 3D PBL parameterization that accounts for these effects.

3.2.1.1 Algorithm Development

i. Model Formulation

Although heterogeneity effects represented through horizontal gradients on a mesoscale grid are likely significantly smaller than vertical gradients, long term effects on the accuracy of wind forecast could be appreciable. We therefore explored implications of eliminating the homogeneity assumption in the development of an improved PBL parameterization. The homogeneity assumption significantly simplifies RANS equations resulting in equations for horizontal components of velocity of the following form

$$\frac{\partial V_1}{\partial t} + V_j \frac{\partial V_1}{\partial x_j} = -\frac{\partial P}{\partial x_1} + \frac{\partial \langle u_1 u_3 \rangle}{\partial x_3}$$

$$\frac{\partial V_2}{\partial t} + V_j \frac{\partial V_2}{\partial x_j} = -\frac{\partial P}{\partial x_2} + \frac{\partial \langle u_2 u_3 \rangle}{\partial x_3}$$

$$\frac{\partial \Theta}{\partial t} + V_j \frac{\partial \Theta}{\partial x_j} = \frac{\partial \langle u_3 \theta \rangle}{\partial x_3}$$

Here, angle brackets denote grid-cell-volume averaging and capital letters denote grid-cell- volume-averaged quantities, V_j denotes velocity components and, Θ potential temperature. Lower case letters denote fluctuating quantities: u_j are fluctuating velocity components, and θ is a fluctuating potential temperature. Repeated indices indicate summation. Under the homogeneity assumption horizontal gradients of turbulent stresses and fluxes are identically zero. This means that only two components of turbulent stress, $\langle u_1 u_3 \rangle$ and $\langle u_2 u_3 \rangle$, and one component of turbulent flux, $\langle u_3 \theta \rangle$, and their vertical gradients affect the evolution of mean fields. Therefore, 1D PBL parameterizations include only these three terms. By eliminating the homogeneity assumption, the prognostic equations for the mean momentum and potential temperature are:

$$\frac{\partial V_i}{\partial t} + V_j \frac{\partial V_i}{\partial x_j} = -\frac{\partial P}{\partial x_i} + \frac{\partial \langle u_i u_j \rangle}{\partial x_j}$$

$$\frac{\partial \Theta}{\partial t} + V_j \frac{\partial \Theta}{\partial x_j} = \frac{\partial \langle u_j \theta \rangle}{\partial x_j}$$

Now all six components of the symmetric turbulent stress tensor need to be parameterized and a full divergence of the stress tensor computed. Similarly, all three components of the turbulent flux must be parameterized and a full divergence of the flux vector computed.

Three-dimensional parameterizations of turbulent stresses and fluxes are commonly used in engineering RANS equations-based model. However, due to large grid cell aspect ratios previously used in NWP models, horizontal gradients in a boundary layer were many orders of magnitude smaller than vertical

gradients. Therefore, the effect of horizontal gradients on boundary layer turbulence development and evolution was neglected. Nevertheless, following development of limited area models in early 1970's (Warner 2010) Mellor and Yamada (1974) outlined a hierarchy of turbulence parameterizations for atmospheric flow simulations that included a fully 3D PBL parameterization. They classified turbulence parameterizations in four levels based on the assumptions made deriving them, with level four representing the full three- dimensional parameterization including prognostic equations for all the second order turbulence moments. In this work we follow the developments of Mellor and Yamada (1982) variant of which were also implemented in the Mellor-Yamada-Nakanishi-Niino (MYNN) PBL parameterization (Nakanishi 2001; Nakanishi and Niino 2004, 2009; Ito et al. 2015). We implemented Mellor- Yamada (Mellor and Yamada 1982) level 2 parameterization that neglects material derivatives of second order moments including the TKE and instead diagnostic equations are provided. First, the TKE is computed using a diagnostic equation

$$\begin{aligned}\frac{q^3}{\Lambda_1} = \; & -\langle u^2\rangle\frac{\partial U}{\partial x} - \langle v^2\rangle\frac{\partial V}{\partial y} - \langle w^2\rangle\frac{\partial W}{\partial z} \\ & -\langle uv\rangle\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right) - \langle uw\rangle\left(\frac{\partial U}{\partial z}+\frac{\partial W}{\partial x}\right) - \langle vw\rangle\left(\frac{\partial V}{\partial z}+\frac{\partial W}{\partial y}\right) \\ & -\beta g\langle w\theta\rangle\end{aligned}$$

Here, q is twice the TKE, Λ_1 is the length scale (sometimes labeled the "dissipation length scale), β is the coefficient of thermal expansion, and g is gravitational acceleration. Once the TKE is diagnosed, second order moments are computed by inverting the following system of linear algebraic equations at each grid cell

$$\begin{bmatrix}
\frac{q}{2\ell_1}+2\frac{\partial U}{\partial x} & -\frac{\partial V}{\partial y} & -\frac{\partial W}{\partial z} & 2\frac{\partial U}{\partial y}-\frac{\partial V}{\partial x} & 2\frac{\partial U}{\partial z}-\frac{\partial W}{\partial x} & -\frac{\partial V}{\partial z}-\frac{\partial W}{\partial y} & 0 & 0 & \beta g & 0 \\
-\frac{\partial U}{\partial x} & \frac{q}{2\ell_1}+2\frac{\partial V}{\partial y} & -\frac{\partial W}{\partial z} & 2\frac{\partial V}{\partial x}-\frac{\partial U}{\partial y} & -\frac{\partial U}{\partial z}-\frac{\partial W}{\partial x} & 2\frac{\partial V}{\partial z}-\frac{\partial W}{\partial y} & 0 & 0 & \beta g & 0 \\
-\frac{\partial U}{\partial x} & -\frac{\partial V}{\partial y} & \frac{q}{2\ell_1}+2\frac{\partial W}{\partial z} & -\frac{\partial U}{\partial y}-\frac{\partial V}{\partial x} & 2\frac{\partial W}{\partial x}-\frac{\partial U}{\partial z} & 2\frac{\partial W}{\partial x}-\frac{\partial V}{\partial z} & 0 & 0 & -2\beta g & 0 \\
\frac{\partial V}{\partial x} & \frac{\partial U}{\partial y} & 0 & \frac{q}{3\ell_1}+\frac{\partial U}{\partial x}+\frac{\partial V}{\partial y} & \frac{\partial V}{\partial z} & \frac{\partial U}{\partial z} & 0 & 0 & 0 & 0 \\
\frac{\partial W}{\partial x} & 0 & \frac{\partial U}{\partial z} & \frac{\partial W}{\partial y} & \frac{q}{3\ell_1}+\frac{\partial U}{\partial x}+\frac{\partial W}{\partial z} & \frac{\partial U}{\partial y} & -\beta g & 0 & 0 & 0 \\
0 & \frac{\partial W}{\partial y} & \frac{\partial V}{\partial z} & \frac{\partial W}{\partial x} & \frac{\partial V}{\partial x} & \frac{q}{3\ell_1}+\frac{\partial V}{\partial y}+\frac{\partial W}{\partial z} & 0 & -\beta g & 0 & 0 \\
\frac{\partial \Theta}{\partial x} & 0 & 0 & \frac{\partial \Theta}{\partial y} & \frac{\partial \Theta}{\partial z} & 0 & \frac{q}{3\ell_2}+\frac{\partial U}{\partial x} & \frac{\partial U}{\partial y} & \frac{\partial U}{\partial z} & 0 \\
0 & \frac{\partial \Theta}{\partial y} & 0 & \frac{\partial \Theta}{\partial x} & 0 & \frac{\partial \Theta}{\partial z} & \frac{\partial V}{\partial x} & \frac{q}{3\ell_2}+\frac{\partial V}{\partial y} & \frac{\partial V}{\partial z} & 0 \\
0 & 0 & \frac{\partial \Theta}{\partial z} & 0 & \frac{\partial \Theta}{\partial x} & \frac{\partial \Theta}{\partial y} & \frac{\partial W}{\partial x} & \frac{\partial W}{\partial y} & \frac{q}{3\ell_2}+\frac{\partial W}{\partial z} & -\beta g \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{\partial \Theta}{\partial x} & \frac{\partial \Theta}{\partial y} & \frac{\partial \Theta}{\partial z} & \frac{q}{\Lambda_2}
\end{bmatrix}
\begin{bmatrix}\overline{u^2}\\ \overline{v^2}\\ \overline{w^2}\\ \overline{uv}\\ \overline{uw}\\ \overline{vw}\\ \overline{u\theta}\\ \overline{v\theta}\\ \overline{w\theta}\\ \overline{\theta^2}\end{bmatrix}
=
\begin{bmatrix}
\frac{q^3}{6\ell_1}+3C_1q^2\frac{\partial U}{\partial x}\\
\frac{q^3}{6\ell_1}+3C_1q^2\frac{\partial V}{\partial y}\\
\frac{q^3}{6\ell_1}+3C_1q^2\frac{\partial W}{\partial z}\\
C_1q^2\left(\frac{\partial V}{\partial x}+\frac{\partial U}{\partial y}\right)\\
C_1q^2\left(\frac{\partial W}{\partial x}+\frac{\partial U}{\partial z}\right)\\
C_1q^2\left(\frac{\partial W}{\partial y}+\frac{\partial V}{\partial z}\right)\\
0\\0\\0\\0
\end{bmatrix}$$

Here, ℓ_1, Λ_1, ℓ_2, and Λ_2 are length scales that are proportional to each other, so they can be expressed in terms of a master length scale ℓ:

$$[\ell_1 \quad \Lambda_1 \quad \ell_2 \quad \Lambda_2] = \ell\,[A_1 \quad B_1 \quad A_2 \quad B_2]$$

The constants, A_1, B_1, A_2, B_2, and C_1 are determined from experimental data. The original values used by Mellor and Yamada (1982) are

$$[A_1 \quad B_1 \quad A_2 \quad B_2 \quad C_1] = [0.92 \quad 16.6 \quad 0.74 \quad 10.1 \quad 0.08]$$

As an intermediate step to implementing a full 3D PBL parameterization we have developed a hybrid approach where all the six components of turbulent stress tensor and three components of the sensible

heat gradient vector are diagnosed and the full divergence of both stress tensor and flux vector computed, but a 1D PBL approximation (i.e. neglecting horizontal derivatives) is used to develop diagnostic equations. This approach leads to the following simplified set of linear algebraic equations

$$\begin{bmatrix} \frac{q}{2\ell_1} & 0 & 0 & 0 & 2\frac{\partial U}{\partial z} & -\frac{\partial V}{\partial z} & 0 & 0 & \beta g & 0 \\ 0 & \frac{q}{2\ell_1} & 0 & 0 & -\frac{\partial U}{\partial z} & 2\frac{\partial V}{\partial z} & 0 & 0 & \beta g & 0 \\ 0 & 0 & \frac{q}{2\ell_1} & 0 & -\frac{\partial U}{\partial z} & -\frac{\partial V}{\partial z} & 0 & 0 & -2\beta g & 0 \\ 0 & 0 & 0 & \frac{q}{3\ell_1} & \frac{\partial V}{\partial z} & \frac{\partial U}{\partial z} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial U}{\partial z} & 0 & \frac{q}{3\ell_1} & 0 & -\beta g & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial V}{\partial z} & 0 & 0 & \frac{q}{3\ell_1} & 0 & -\beta g & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\partial \Theta}{\partial z} & 0 & \frac{q}{3\ell_2} & 0 & \frac{\partial U}{\partial z} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\partial \Theta}{\partial z} & 0 & \frac{q}{3\ell_2} & \frac{\partial V}{\partial z} & 0 \\ 0 & 0 & \frac{\partial \Theta}{\partial z} & 0 & 0 & 0 & 0 & 0 & \frac{q}{3\ell_2} & -\beta g \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\partial \Theta}{\partial z} & \frac{q}{\Lambda_2} \end{bmatrix} \begin{bmatrix} \overline{u^2} \\ \overline{v^2} \\ \overline{w^2} \\ \overline{uv} \\ \overline{uw} \\ \overline{vw} \\ \overline{u\theta} \\ \overline{v\theta} \\ \overline{w\theta} \\ \overline{\theta^2} \end{bmatrix} = \begin{bmatrix} \frac{q^3}{6\ell_1} \\ \frac{q^3}{6\ell_1} \\ \frac{q^3}{6\ell_1} \\ 0 \\ C_1 q^2 \frac{\partial U}{\partial z} \\ C_1 q^2 \frac{\partial V}{\partial z} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

ii. Model Implementation

We implemented the new 3D PBL parameterization in the WRF-ARW NWP model. Unlike 1D PBL parameterizations implemented in WRF-ARW that are grouped together with other vertical column-based parameterizations of physical processes in subdirectory /phys the 3D PBL parameterization we included in the /dyn_em subdirectory. All the subroutines directly related to the 3D PBL parameterization are collected in the Fortran module module_turb_mixing_3d.F. Additional changes include introduction of state variables and namelist configuration parameters associated with the 3D PBL parameterization in the registry file Registry.EM_COMMON. The WRF-ARW source code with the 3D PBL parameterization has been stored in a github repository at: https://github.com/NCAR/WFIP2-WRF-3DPBL/blob/master/wrf3.8.1.

iii. Mesoscale Simulations over Idealized Heterogeneous Terrain

The new PBL scheme is able to account for horizontal heterogeneity of boundary-layer flows not only due to complex topography and heterogenous land use, but also due to large convective eddies, convective cells and rolls. Convective rolls and cells generally scale with the boundary layer height and therefore, as the grid cell size approaches boundary-layer height, these structures are being captured by NWP models, however, they are under-resolved (Ching et al. 2014). Since the structures are under-resolve their characteristic length scales are not realistic. It is therefore important to account for the effects of velocity and potential temperature gradients induced by convective structures that result in enhanced turbulence diffusion. Other convectively induced secondary circulations caused by, for example, heterogeneous surface characteristics are often not resolved properly using NWP models with 1D PBL parameterizations and grid cell size in “terra incognita” (i.e. gray zone) range.

Before proceeding with simulations over heterogeneous terrain, model parameters were tuned using simulations of a convective boundary layer over homogeneous terrain. We found that when using parameters determined by Bougeault and Lacarrere (1989) the 3D PBL parameterization produces the correct mixed layer ABL structure while the original set of parameters presented by Mellor and Yamada (1982) result in a diabatic profile of potential temperature (Martilli, personal communication). Therefore, the parameters that we used in all the simulations presented here are:

$$[A_1 \quad B_1 \quad A_2 \quad B_2 \quad C_1] = [0.3 \quad 8.4 \quad 0.33 \quad 6.4 \quad 0.08]$$

Not accounting for the effects of horizontal gradients on turbulence production, transport, and dissipation results in inaccurate levels of turbulence and therefore inaccurate turbulent diffusion. We demonstrated

this by carrying out idealized mesoscale simulations of an ABL over heterogeneous surface. The results of these simulations are verified by comparing them to the results of an ensemble of LES. The idealized mesoscale simulation setup includes periodic lateral boundary conditions with a weak, 2 ms^{-1}, southerly wind and surface heat flux of 160 Wm^{-2} on the west half of the domain and 320 Wm^{-2} on the east half of the domain. The horizontal grid cell size in both directions was 200 m. Two mesoscale simulations were carried out, one using the MYNN 1D PBL parameterization and the other using the new 3D PBL parameterization.

Additionally, an ensemble of 20 LES was carried out over the lateral boundary conditions and the same domain but with grid cell size of 50 m. The ensemble was created by adding a uniformly distributed random perturbation to the surface heat flux. All the simulations were run for two hours of physical time. The results of these simulations are shown in Figure 5 and Figure 6. In these figures shown are contour plots of meridional and vertical velocity components at 250 m above the surface, respectively. In the left panel of Figure 5 presented are mesoscale simulation results obtained using the 1D PBL parameterization while on the right are results from the simulation with the 3D PBL parameterization. In Figure 6, the left and middle panel show mesoscale simulation results with the 1D and 3D PBL parameterizations, respectively, while the right panel shows the ensemble average vertical velocity from the LES.

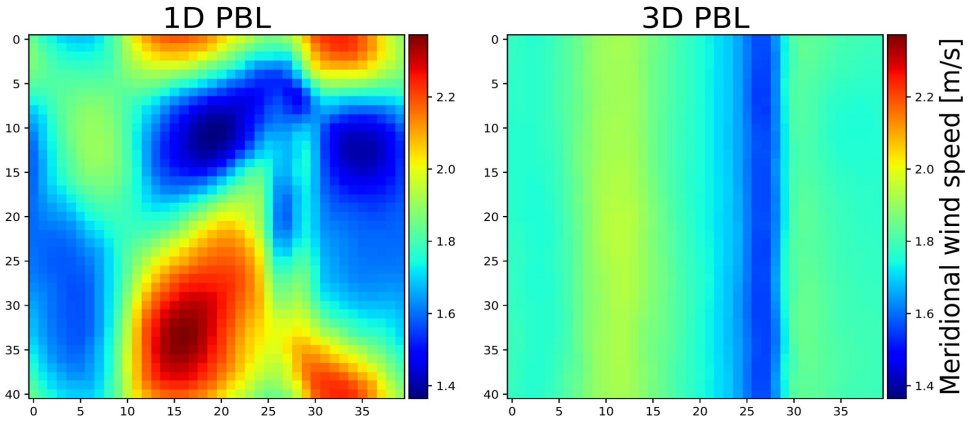

Figure 5. Meridional velocity from mesoscale simulations with the 1D PBL parameterization (left) and 3d PBL parameterization (right).

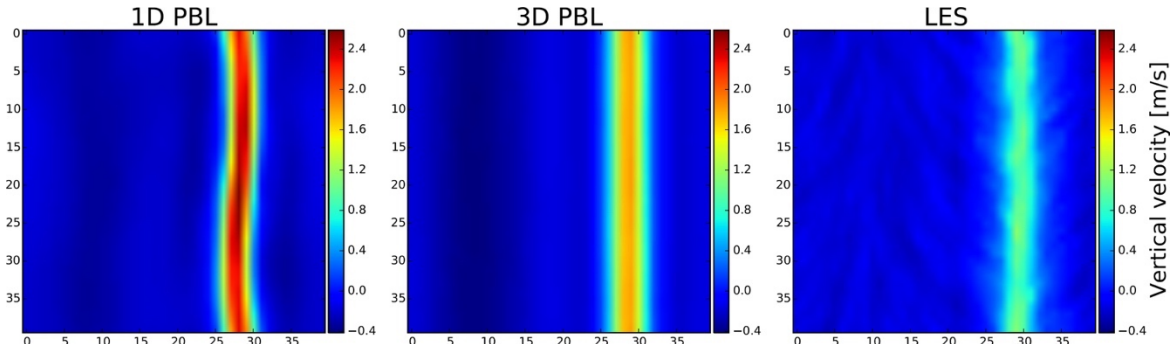

Figure 6. Vertical velocity contours from mesoscale simulations with the 1D PBL parameterization (left), the 3D PBL parameterization (middle), and the ensemble average LES (right).

It is clear from Figure 5 that using the 1D PBL parameterization to simulate a heterogeneous convective ABL at 200 m results in a significant unphysical meridional (i.e. streamwise) variability in the meridional velocity component. Less pronounced, but still clearly evident, is an unphysical variability in the vertical velocity shown in Figure 6. The 3D PBL parameterization, on the other hand, results in the vertical velocity filed that is homogeneous in the meridional direction and similar to the ensemble mean of the LES. We can conclude that favorable comparison of the 3D PBL simulation results to the LES ensemble results verifies that the 3D PBL parameterization better represents the effects of the horizontal turbulent diffusion due to horizontal gradients of velocity and potential temperature.

iv. Large-Eddy Simulations over Complex Terrain

The data collected during the filed study enabled our team to attempt to improve the understanding of the fundamental physical processes and their representation in forecast models, thereby improving forecast capabilities. While the WFIP2 field study data represent an invaluable resource for studying flows in complex terrain, majority of the observations were in situ observations relatively sparsely distributed over a very large field study area. A few scanning lidars provided spatial information, however, their limited range could not fully capture the complexity of the flow West of the Columbia River Gorge. In the attempt to supplement observations and provide more information about spatial structures of the flow, including mountain waves, topographic wakes, and gap flows, we have carried out high-resolution simulations, using an LES approach, of a selected observational period when all these phenomena were observed. With the LES approach we were able to resolved the details of the flow. The validated LES results can then be used in conjunction with observations to more completely validate the mesoscale simulations with the new 3D PBL parameterization.

We used nesting capability in the WRF-ARW to nest LES within a mesoscale domain. The outer, parent domain and the inner, nested domain are shown in Figure 7 and Figure 8. The parent domain spans 1800 km in west to east direction and 900 km in South to North direction. The inner, LES domain spans 180 km in west to east direction and 90 km south to north direction. Both domains are resolved with 6000 grid cells in west to east direction and 3000 grid cells in south to north directions. While the parent domain grid cell size is 300 m while the inner domain grid cell size is 30 m. As can be seen from Figure 7, the outer domain spans from the Pacific Ocean on the west to high plains of Montana and Wyoming and encompasses northern Oregon and southern Washington states. The inner domain is centered on the area south of the Columbia River where there is a large concentration of deployed wind power capacity. It includes Cascades and Mountain Hood, as well as Columbia River Gorge to the east. The grey scale indicates elevation with darker colors representing higher elevation. The dark spot in the lower left area of the plot represents Mountain Hood. Since the prevalent winds are from the west, complex terrain features affect the flow patterns and wind resource to the west where wind plants are located.

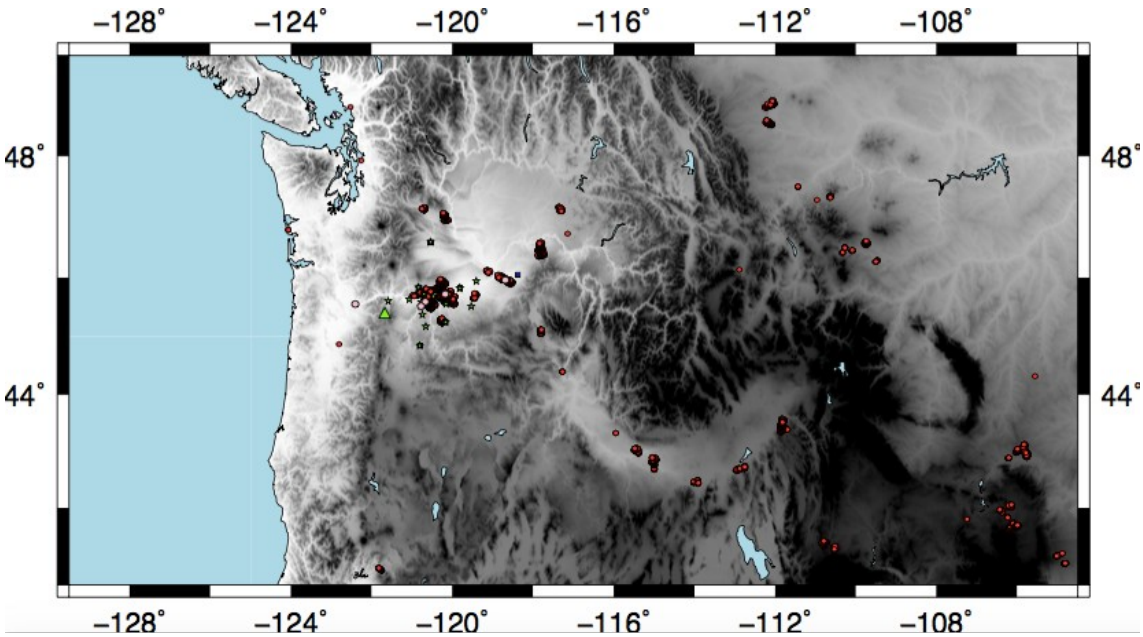

Figure 7. Outer, parent mesoscale domain for nested simulations over the WFIP2 filed study area.

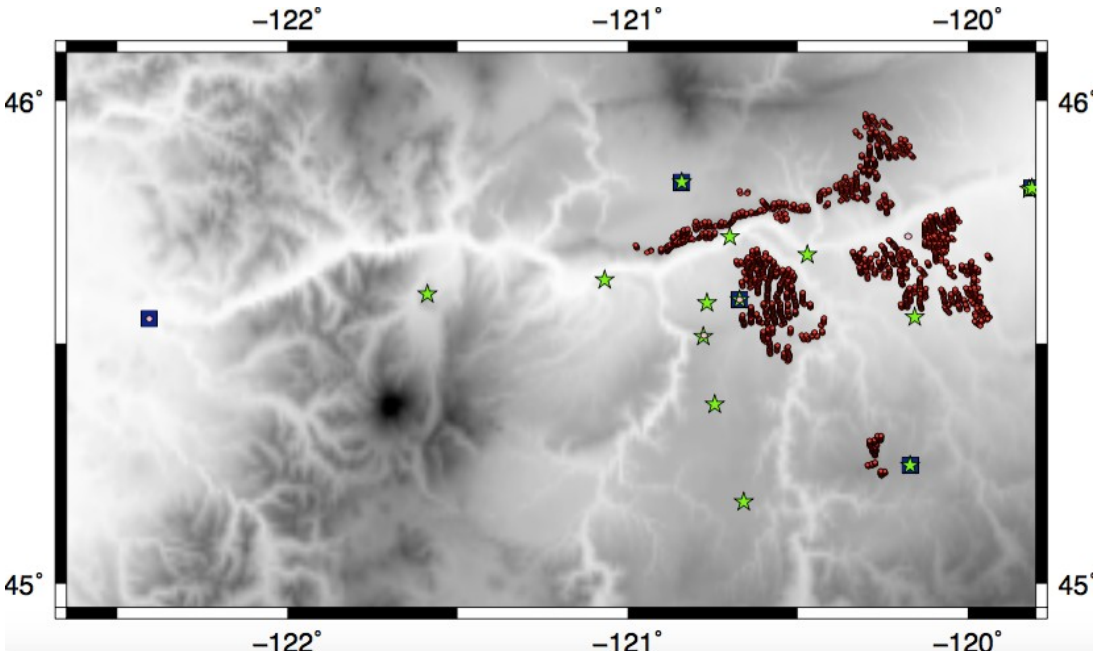

Figure 8. Inner, nested LES domain for nested simulations over the WFIP2 filed study area.

To study the complex terrain effects on the flow patterns we have selected to simulate weather conditions observed on March 7 and 8. Our simulations spanned 12 hours, from 15 UTC on March 7 to 3 UTC on March 8. The weather conditions on March 7 and 8 were characterized with steady westerly flows resulting in an accelerated gap flow through the Columbia River Gorge that spilled east of the Gorge. The orography including the Cascades Range and Mountain Hood generated mountain waves.

Mountain Hood, shown in Figure 9, is a volcanic mountain rising to 3500 m above sea level (ASL), or close to 3 km above the surrounding area. It is about 10 km in diameter and therefore it represents a distinct, isolated obstacle to the westerly flow, causing in a well-defined topographic wake.

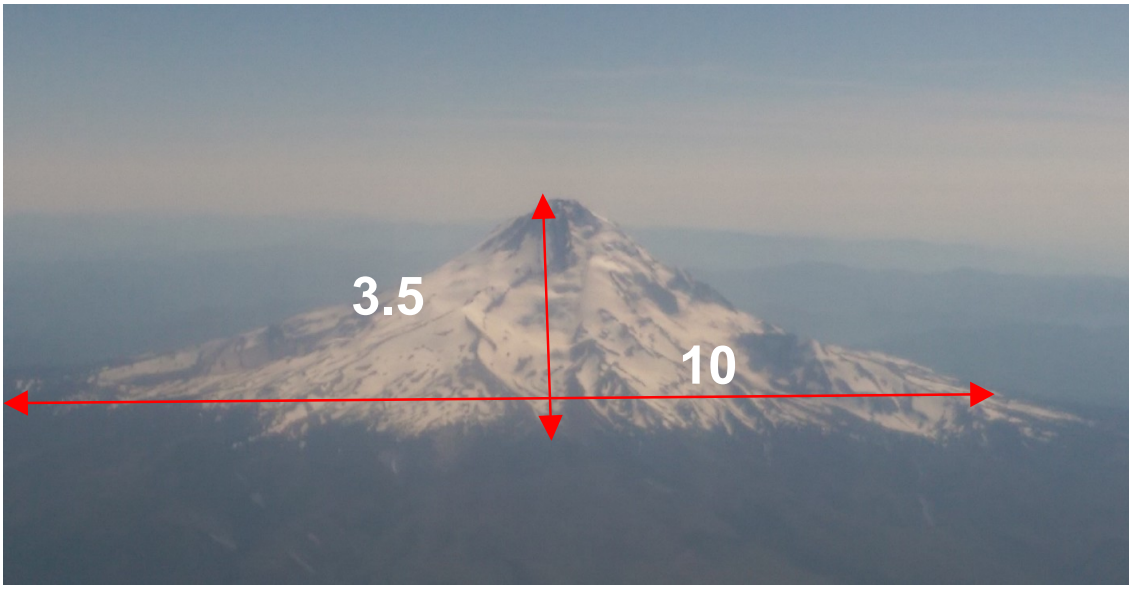

Figure 9. Mount Hood

In Figure 10 is shown the evolution of vertical profile of potential temperature observed using a radiometer located at the Wasco airport. A capping inversion can be identified at about 1900 m due to a steep potential temperature gradient. From 15 UTC on March 7 to 1 UTC on March 8 an adiabatic layer can be observed below 500 m with a well-mixed layer above it signifying a daytime convective boundary layer. These observations indicate that the capping inversion is significantly below the Mountain Hood peek. Under such conditions we can expect that the dividing streamline for the westerly flow is also below the Mountain Hood peak, causing the flow to split around the mountain, creating a orographic wake downstream with potential formation and shedding of von Karman vortices.

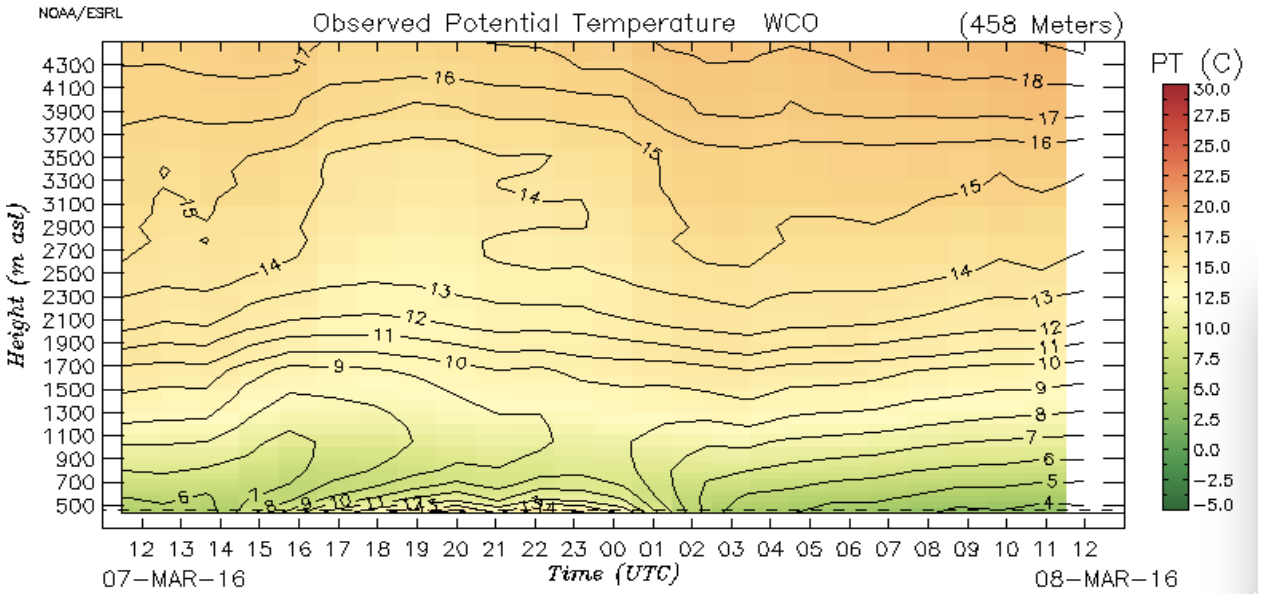

Figure 10. Profiles of potential temperature observed with the University of Colorado radiometer at the Wasco airport on March 7 and 8.

Our nested LES capture both the orographic wake (Figure 11) and von Karman vortices shedding off of Mountain Hood. Both of these flow features can be clearly observed in the animations of LES simulations that can be found at: https:// https://www.seedme.org/node/169835, https://www.seedme.org/node/169348, https://www.seedme.org/node/170118, and https://www.seedme.org/node/170607.

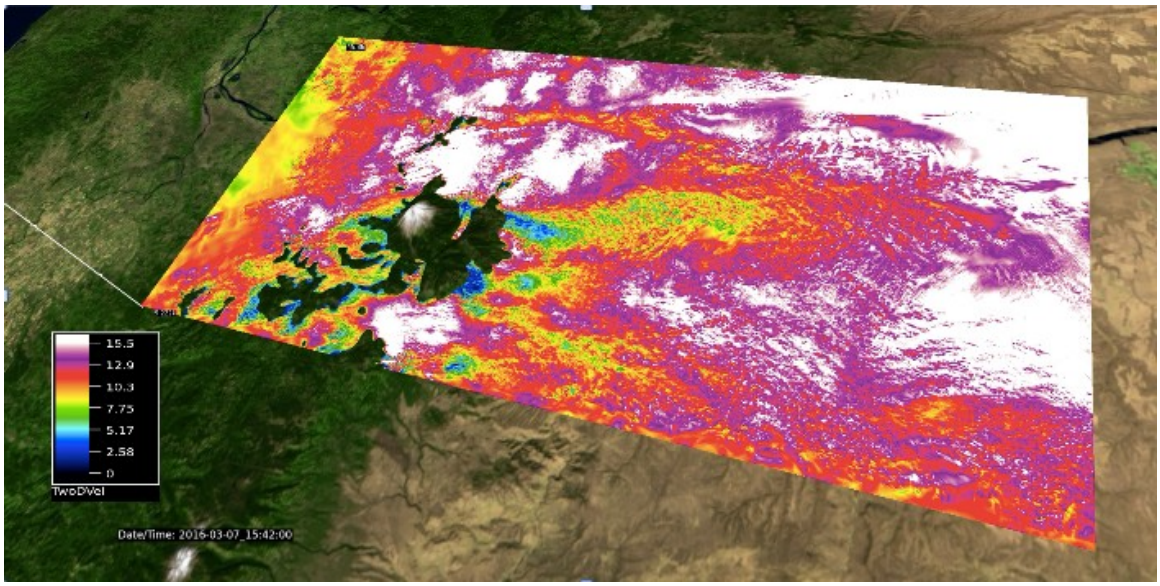

Figure 11. A contour plot of horizontal wind speed from an LES of March 7 and 8 at the Columbia River Gorge.

However, von Karman vortices do not take the form of a regular vortex streets observed in laboratory experiments or in the cloud patterns in wakes of islands. Due to the presence of a ridge to the east of Mountain Hood the regular vortex shedding is disrupted. The LES also captures standing and travelling waves created by the Cascades, Mountain Hood and the ridge to the east of Mountain Hood, as can be seen from Figure 12 and the following animations: https://www.seedme.org/node/170851 and https://www.seedme.org/node/171095.

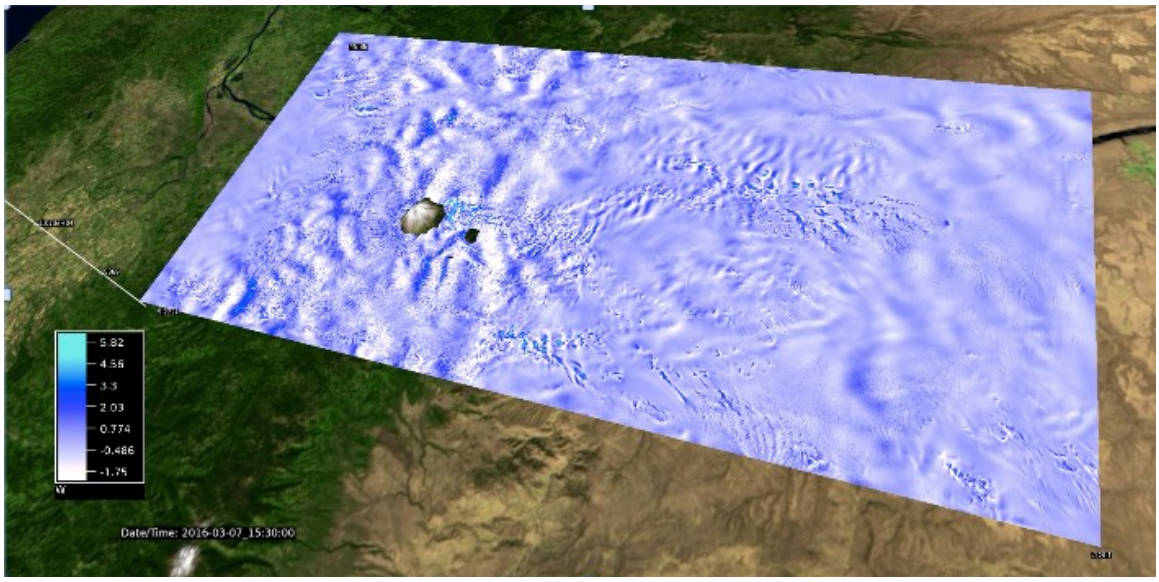

Figure 12. A contour plot of vertical wind speed from an LES of March 7 and 8 at the Columbia River Gorge.

Qualitative analysis of flow patterns revealed by LES animations confirm expectations and casual observations. However, to validate LES a quantitative analysis must be carried out. For that purpose, we use continuous observations of wind speed at BPA towers located in the Columbia River Gorge region as well as observations by Vaisala sodars. The mesoscale results are provided by the parent domain, while LES results are obtained from several simulations carried out to demonstrate the effect of a boundary perturbation as well as a hybrid advection scheme. The baseline LES simulation uses the Smagorinsky subgrid turbulence parameterization, a fifth-order, upwind advection scheme in horizontal direction and a third-order, upwind scheme in vertical direction. The second LES uses the boundary perturbation of potential temperature to speed up turbulence development, while the third LES, in addition, uses the hybrid advection scheme. When LES is nested into a mesoscale domain the inflow to LES domain is smooth and it does not include any resolved three-dimensional turbulent eddies. Due to the smooth inflow conditions a long fetch is needed for turbulence to develop under neutrally stratified conditions over a flat terrain. Turbulence develops significantly faster under convective conditions that characterize most of the LES performed here. In addition, complex terrain contributes to turbulence development.

Nevertheless, we introduced the boundary perturbation of potential temperature following Muñoz-Esparza et al. (2014, 2015) to enhance turbulence initiation. To further improve turbulence development and resolution we have also used the hybrid advection scheme that combines odd order and even order advection schemes. An odd order upwind advection scheme is dissipative resulting in an effective resolution of approximately 7 Δx (Skamarock 2004). This can be compared to a common implementation of a pseudo spectral scheme which has an effective resolution of 3 Δx. By combining odd and even schemes an effective resolution between 4 and 5 Δx can be achieved.

Comparisons were made between LES and and observations taken by BPA towers and Vaisala sodars. Locations of the observations are shown in Figure 13. Examples of comparisons are given in Figure 14 and Figure 15. In these figures are compared the simulation mean absolute errors (MEA) of mesoscale and large-eddy simulations computed based on the wind speed observations at the BPA towers and the Vaisala sodars, respectively. It can be observed that the LES most of the time follow closely mesoscale simulation, however, there are periods when they diverge from the mesoscale wind speed predictions. For example, between UTC 0 and 2 wind speeds from LES at both the Hood River tower and the Vaisala sodar, AON5, are deviating from the mesoscale simulation wind speeds. However, they are in better agreement with observations than mesoscale simulations. This can be observed at other locations where observations are made. In general, LES result in lower MAE and RMSE as can be seen from Table 5 through Table 8. More significant exception is the Roosevelt tower location on the north bank of the Columbia River.

Since both MAE and RMSE are relatively low at most observation locations and since LES provides an improvement in comparison to the mesoscale simulation results, we can conclude that LES can be used as an additional source of information about the flow in complex terrain. In particular, LES can be used to characterize the horizontal heterogeneity of the flow in the WFIP2 field study area. The LES results can be used in the future to guide further development of the 3D PBL parameterization including calibration of different model parameters.

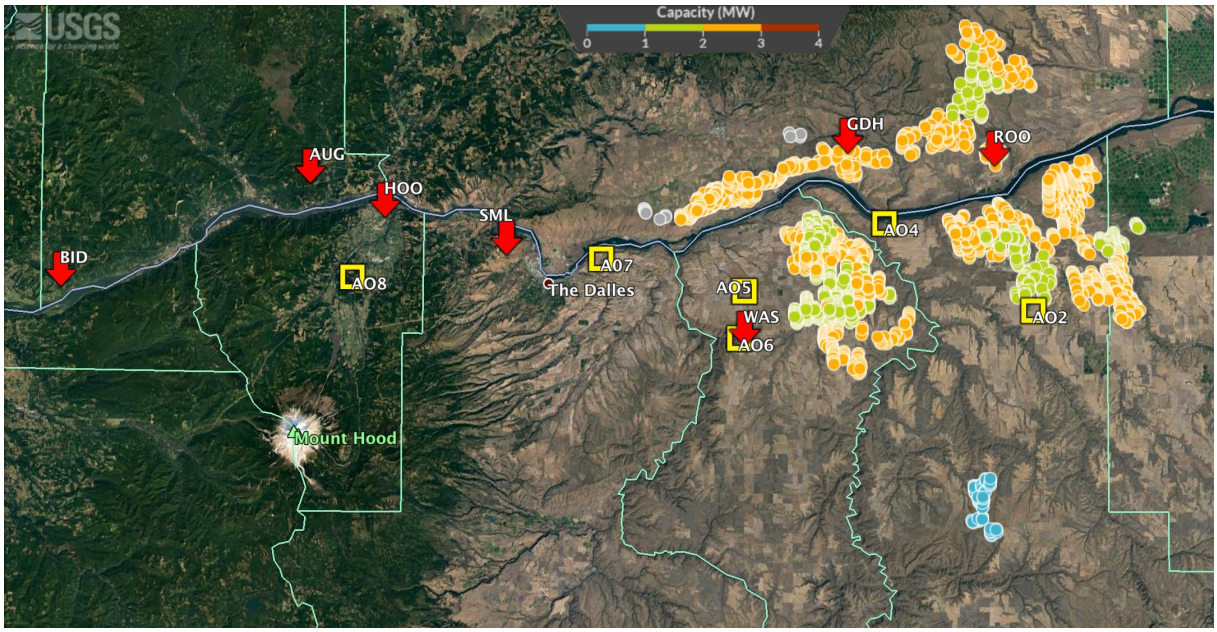

Figure 13. Location of BPA towers (red arrows), Vaisala sodars (yellow squares) and turbines (yellow and green dots) in the Columbia River Gorge field study area.

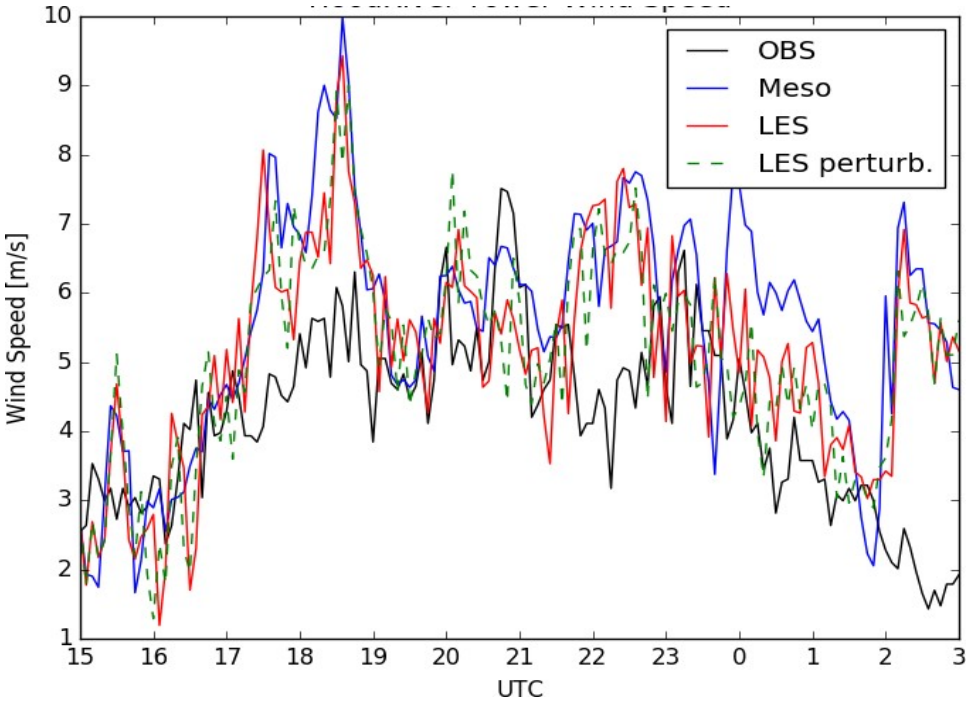

Figure 14. Observed and simulated wind speed at the Hood River BPA tower: observed wind speeds - solid black line, mesoscale simulations – solid black line, LES – solid red line, LES with perturbations – dashed green line.

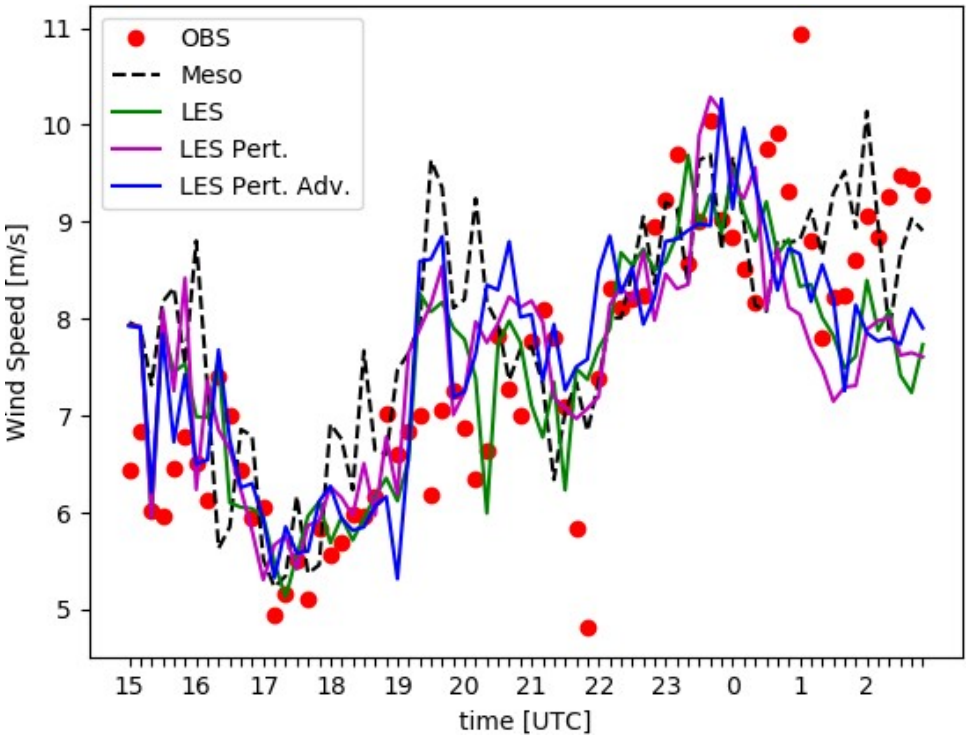

Figure 15. Observed and simulated wind speed at the AON5 sodar location: observations – red dots, mesoscale simulation – dashed black line, LES – solid green line, LES with perturbations – solid purple line, LES with perturbations and a hybrid advection scheme –blue line.

Table 5. MAE of mesoscale and large-eddy simulations based on BPA tower observations.

Mean Absolute Error [m/s]			
	Mesoscale	LES	LES with cell perturbation
Augspurger	4.73	3.99	3.91
Biddle Butte	2.71	2.39	2.43
Goodnoe Hills	1.53	1.16	1.18
Hood River	1.59	1.39	1.30
Roosevelt	1.81	2.02	2.03
Seven Mile Hill	1.74	1.82	1.79
Wasco	2.53	2.44	2.52

Table 6. RMSE of mesoscale and large-eddy simulations based on BPA tower observations.

Root Mean Square Error [m/s]			
	Mesoscale	LES	LES with cell perturbation
Augspurger	5.11	4.42	4.35
Biddle Butte	3.48	3.12	3.09
Goodnoe Hills	1.94	1.46	1.51
Hood River	2.04	1.76	1.68
Roosevelt	2.13	2.36	2.40
Seven Mile Hill	2.14	2.21	2.18
Wasco	3.25	3.18	3.22

Table 7. MAE of mesoscale and large-eddy simulations based on Vaisala sodar observations.

Mean Absolute Error [m/s]				
	Mesoscale	LES	LES with cell perturbation	LES with cell perturbation and hybrid advection
AON2	2.56	2.26	2.31	2.36
AON4	1.63	1.14	1.18	1.15
AON5	0.89	0.75	0.84	0.79
AON6	2.45	2.47	2.60	2.45
AON7	2.14	2.20	2.16	2.24
AON8	2.70	2.73	2.63	2.63

Table 8. RMSE of mesoscale and large-eddy simulations based on Vaisala sodar observations.

Root Mean Square Error [m/s]				
	Mesoscale	LES	LES with cell perturbation	LES with cell perturbation and hybrid advection
AON2	3.21	2.76	2.83	2.90
AON4	2.02	1.40	1.49	1.45
AON5	1.15	0.96	1.04	1.02
AON6	3.26	3.26	3.37	3.32
AON7	2.71	2.80	2.76	2.91
AON8	3.47	3.52	3.44	3.42

v. Mesoscale Simulations over Complex Terrain

We validated WRF-ARW with a new 3D PBL parameterization using field experiment data from the Columbia River Gorge. For this purpose, we focused on validation of the parameterization for all the components of turbulent stress and sensible heat flux and their full divergence, but with the use of a boundary-layer approximation (see Appendix). This approach was assessed first rather than the full 3D parameterization due to numerical stability issues related to the use of diagnostic equation for the TKE that can result in ill conditioning of the system of linear algebraic equations.

The validation of 3D PBL parameterization is based on one of the selected “ten-day” WFIP2 retrospective study periods, the period between August 13 and 24, 2016. We carried out three mesoscale simulations. We used the MYNN 1D PBL parameterization for the baseline mesoscale simulation. In the other two mesoscale simulations, we used the new 3D PBL with and without an additional two-dimensional (2D) Smagorinsky type diffusion parameterization commonly used in operational NWP models. Here, it should be pointed out that the role of the 2D Smagorinsky diffusion parameterization is not to represent unresolved physical processes and resulting diffusion (Skamarock 2004; Smagorinsky 1990), but instead to provide numerical stability through diffusion of numerical oscillations controlled by a strain-rate-dependent diffusivity. The initial and boundary conditions were derived from HRRR simulations with horizontal grid cell size of 3 km. The HRRR output was obtained from the University of Utah data archive (http://home.chpc.utah.edu/~u0553130/Brian_Blaylock/hrrr.html) maintained by Brian Blaylock. Simulations were carried out for 30 hours. The first 6 hours represented simulation spin up time and the last six hours overlapped with the next day’s simulation. All the simulations started at 00 UTC and ended the next day at 06 UTC. The simulations were carried out using two domains. The parent domain was resolved using 750 m horizontal grid cell sizes, while the inner, nested domain was resolved with 250 m horizontal grid cell sizes (Figure 16).

The inner domain was centered on the so-called Physics Site near just east of the Biglow Canyon wind plant. The outer domain resolution corresponded to the resolution of a special instance of HRRR that was run by the NOAA team over the WFIP2 study area during the duration of the field study.

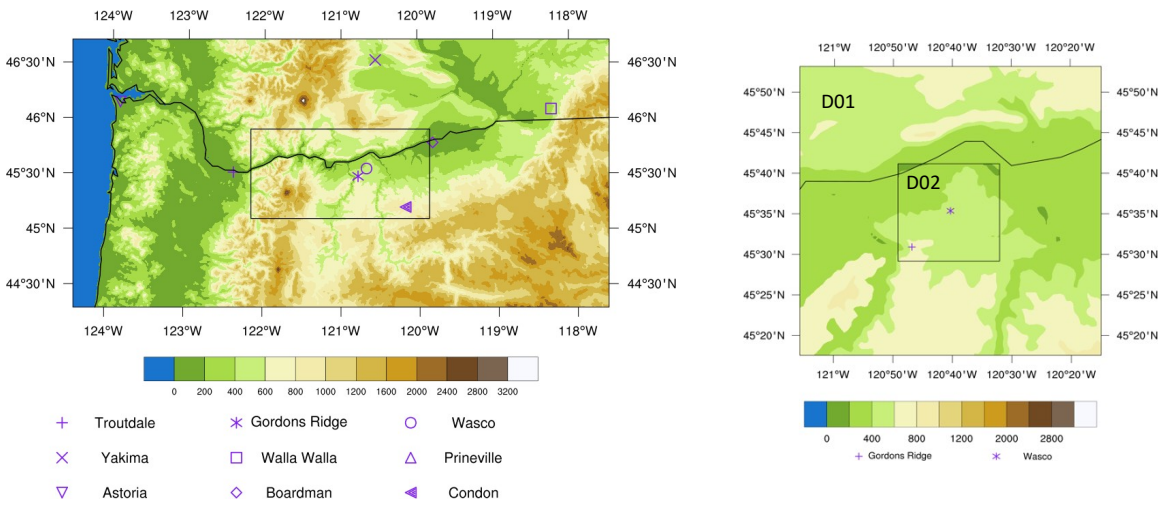

Figure 16. Left panel - WFIP2 field study area with symbols indicating instrument locations; Right panel – WRF domains, D01 – 750 m grid cell size, D02 – 250 m grid cell size. In DO2 plus sign denotes Wasco tower and x denotes Physics site (by Masih Eghdami).

The observations used to assess the model performance are from observational platforms located within the inner domain including: two-meter tower at Wasco – wind only, three-meter tower at the Physics site: wind data, relative humidity, temperature, and irradiance, and a 17- meter Physics site tower with sonic measurements at three levels: 3 m, 10 m, and 17 m. Sonics provide high-frequency measurements needed to compute turbulent fluxes.

Shown in Figure 17 are comparisons of output from the three mesoscale simulations with observations at the two-meter tower at Wasco. In this figure shown are example comparisons based on observations from two simulated days. From Figure 17 we can see that all the mesoscale simulations capture quite well the diurnal temperature evolution. During nighttime between hours 12 and 19 UTC temperature is first only slightly underpredicted and then during the second day slightly overpredicted.

The kinematic sensible heat flux is shown in the top right panel of Figure 17. In this case the sensible heat flux from simulation with the MYNN parameterization was not included in the output and therefore it is zero. For the other two simulations the agreement between observations and model results is relatively good, however, during both days daytime sensible heat flux is underpredicted. Finally, the wind speed is predicted well except during the first few hours on the first day.

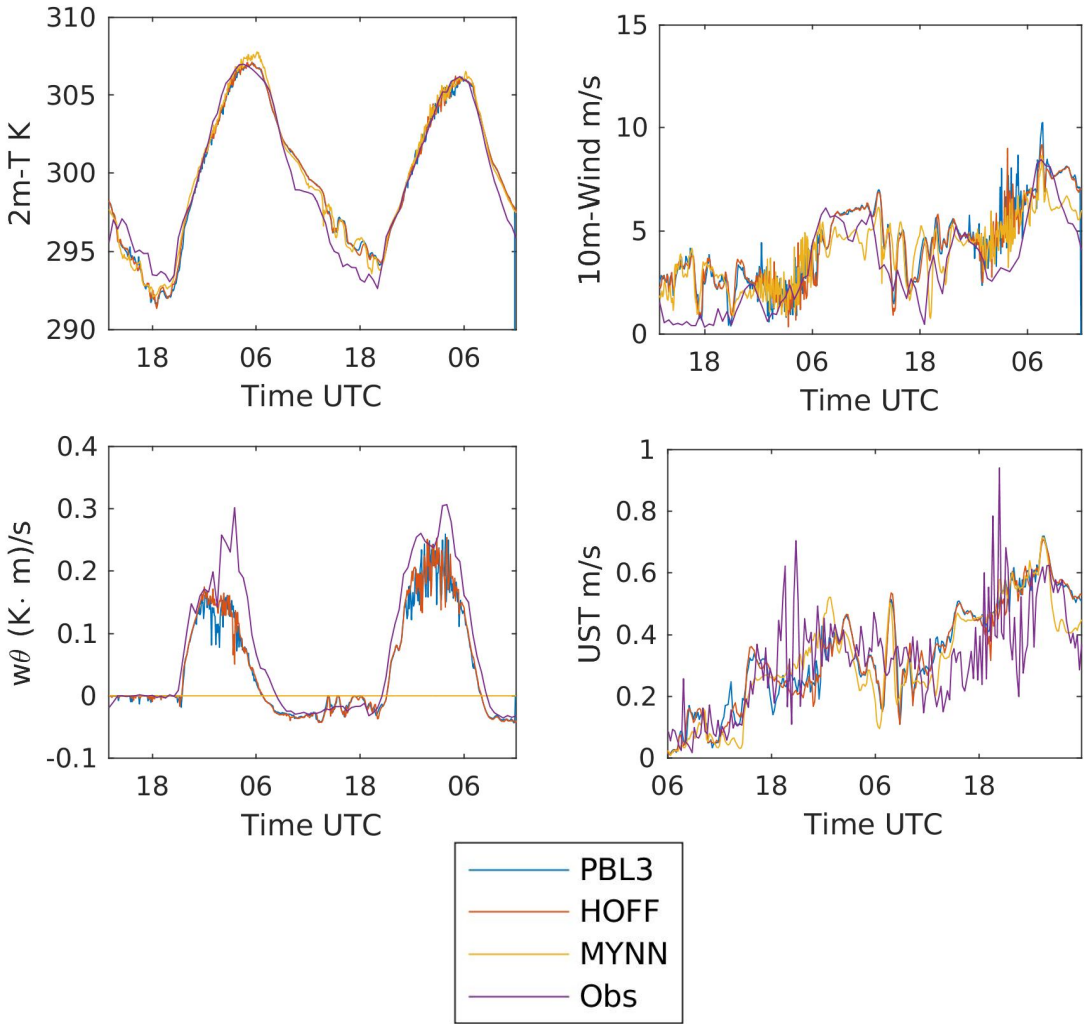

Figure 17. Comparison of the output from domain 2 of three mesoscale simulations with observations at two-meter tower at Wasco: 2 m temperature – top left panel, wind speed at 10 m – top right panel, surface kinematic sensible heat flux – bottom left panel, surface friction velocity – bottom right panel. In all the panels different lines represent: simulation with the 3D PBL parameterization – blue line, simulation with the 3D PBL but with the 2D Smagorinsky parameterization turned off – orange line, simulation with MYNN PBL parameterization – yellow line, and observations – purple line (by Masih Eghdami).

One of the reasons to replace a 1D PBL parameterization with a 3D PBL parameterization is the ability of the former to represent normal turbulent stresses as well as their effects on the momentum evolution. In Figure 18 shown is comparison of the normal turbulent stress components from mesoscale simulations to the observed normal stress components. The horizontal normal stress components are underpredicted by the 3D PBL parameterization while the vertical component is overpredicted. The observations indicate that at the scale of interest normal stresses are significant. Their temporal variability, and therefore likely spatial variability is also significant.

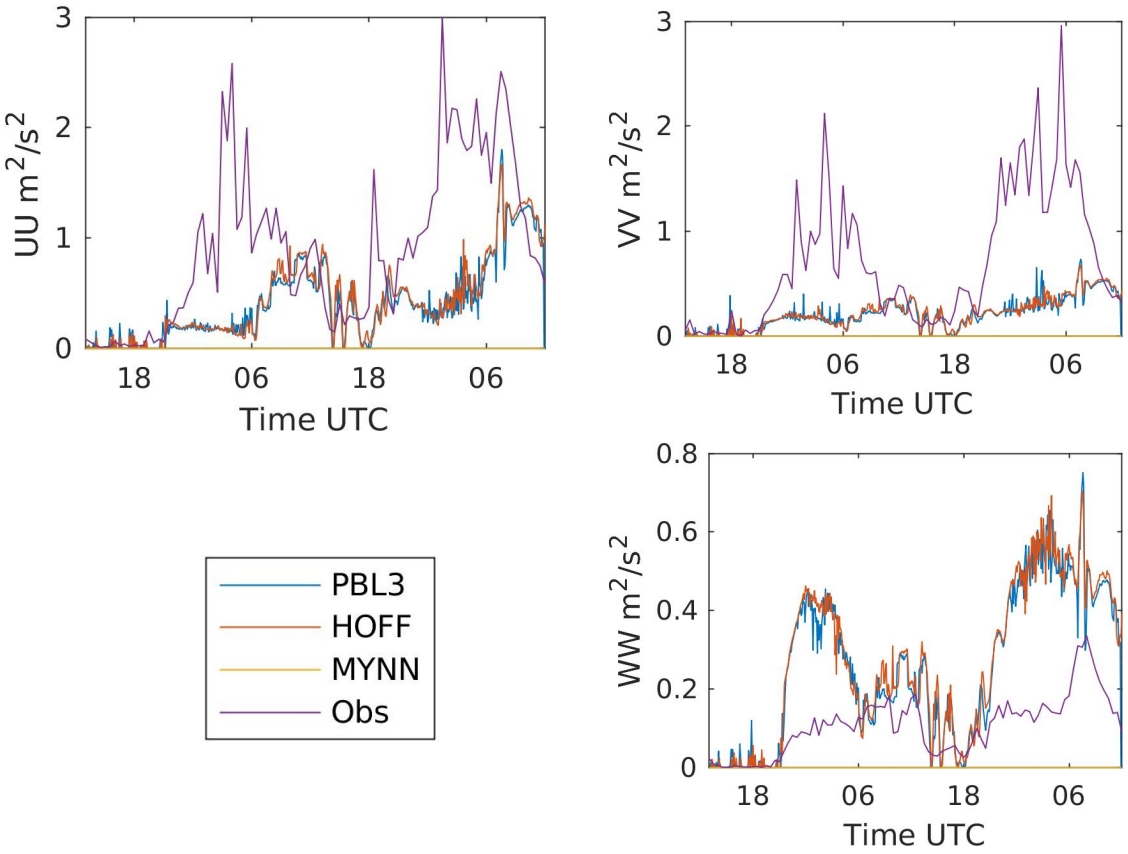

Figure 18. Comparison of simulated and observed normal turbulent stresses: normal turbulent stress component in zonal direction – top left panel; normal turbulent stress component in meridional direction – top right panel; normal turbulent stress component in vertical direction -- bottom right panel. In all the panels different lines represent: simulation with the 3D PBL parameterization – blue line, simulation with the 3D PBL but with the 2D Smagorinsky parameterization turned off – orange line, simulation with MYNN PBL parameterization – yellow line, and observations – purple line (by Masih Eghdami).

These observations point to a need to represent a full turbulent stress tensor and its divergence in high-resolution mesoscale simulations in complex terrain. While 1D PBL parameterizations represent two components of the turbulent shear stress, $\langle u'w' \rangle$ and $\langle v'w' \rangle$, they do not represent the third turbulent shear stress component, $\langle u'v' \rangle$. Again, turbulent shear stresses from mesoscale simulation with MYNN model were not output and therefore in Figure 19 they are all zero. In Figure 19 we observe that both mesoscale simulations including the 3D PBL parameterization predict the magnitude and variation of the two turbulent shear stress components, $\langle u'w' \rangle$ and $\langle v'w' \rangle$, quite well. However, the third turbulent shear stress component, $\langle u'v' \rangle$ is underpredicted.

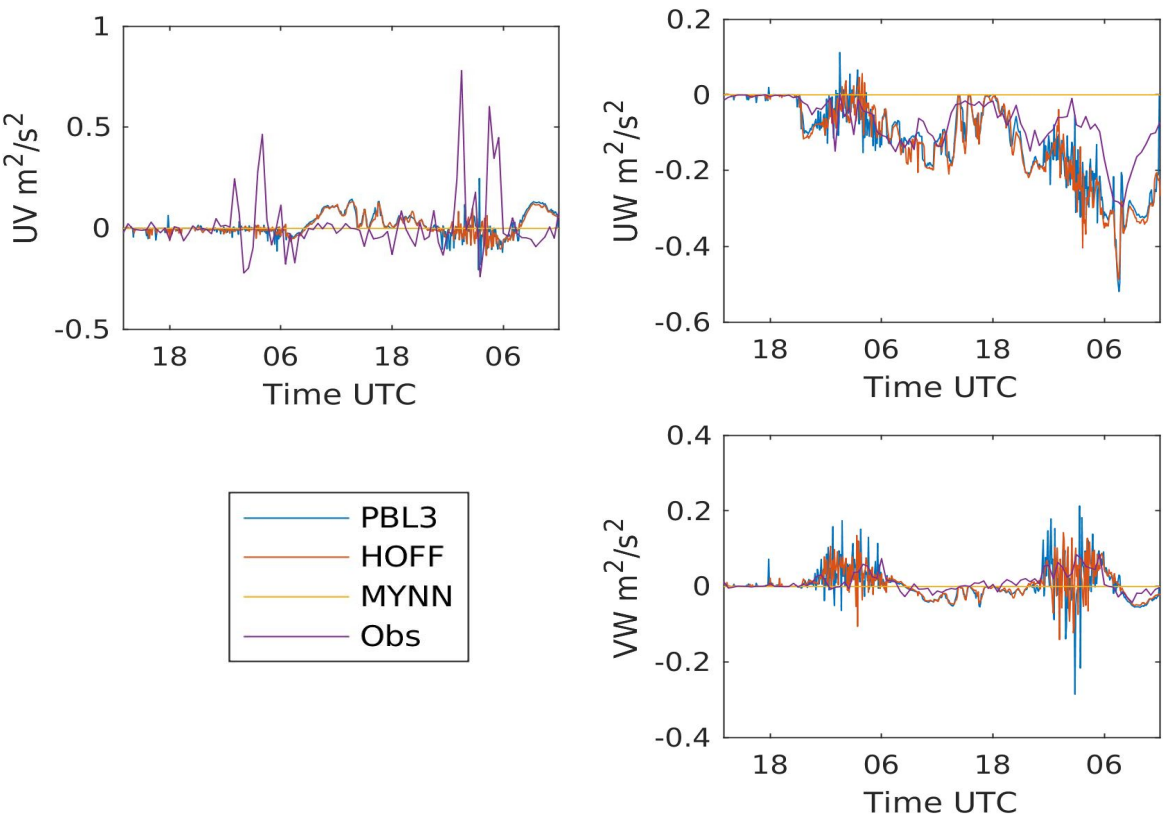

Figure 19. Comparison of simulated and observed turbulent shear stresses: turbulent shear stress component $\langle u'v' \rangle$ – top left panel; turbulent shear stress component $\langle u'w' \rangle$ – top right panel; turbulent shear stress component $\langle v'w' \rangle$– bottom right panel. In all the panels different lines represent: simulation with the 3D PBL parameterization – blue line, simulation with the 3D PBL but with the 2D Smagorinsky numerical diffusion turned off – orange line, simulation with MYNN PBL parameterization – yellow line, and observations – purple line (by Masih Eghdami).

A graphical summary of the mesoscale simulation results is given in Figure 20 in the form of a Taylor diagram. A Taylor diagram combines three statistical measures quantifying the degree of correspondence between observations and model output: the Pearson correlation coefficient, the RMES, and the standard deviation. Review of the Taylor diagram shown in Figure 20 reveals that the use of the 3D PBL parameterization slightly improves correlation with observations compared to simulations with the 1D PBL parameterization. Using the 3D PBL parameterization significantly improves prediction of variability as indicated by the standard deviation being closer to observed than when the 1D PBL parameterization is used. In general simulations with the 3D PBL parameterizations perform similarly regardless of whether the 2D Smagorinsky numerical diffusion is turned on or not.

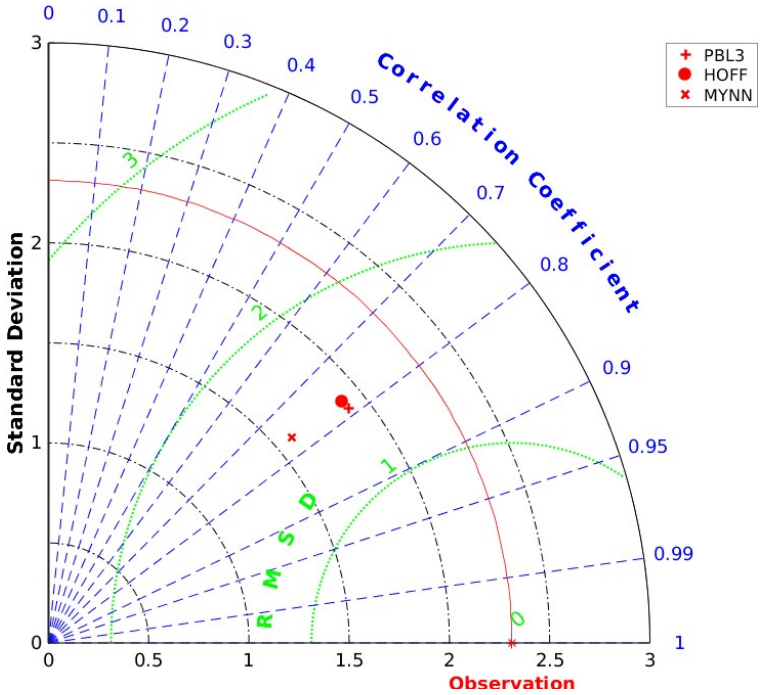

Figure 20. Taylor diagram summary of mesoscale simulation results corresponding to the "ten-day" retrospective period from August 13-24, 2016. Three mesoscale simulations use: 1D PBL MYNN parameterization – x symbol, 3D PBL parameterization – cross, and 3D PBL ameterization with 2D Smagorinsky numerical diffusion turned off (by Masih Eghdami).

3.2.1.2 Summary and Future Work

At Present most NWP models include parameterizations of turbulent stresses and fluxes based on the assumption of horizontal homogeneity over a grid cell and therefore reduced to 1D PBL parameterizations. As the horizontal grid cell sizes decrease the assumption of horizontal homogeneity is violated and the effects of neglected terms must be accounted for. Under convective atmospheric conditions the homogeneity assumption is violated even in flows over flat, homogeneous surface due to the presence of large convectively induced secondary circulations (Ching et al. 2014). We have therefore developed a 3D PBL parameterization following the work of Mellor and Yamada (1974, 1982) and implemented it in the WRF model. The new 3D PBL parameterization was first assessed by carrying out idealized mesoscale simulations over heterogeneous terrain characterized by sharp differences in surface heat fluxes. These simulations demonstrated the deficiency of a 1D PBL parameterization when grid cell size is in the so called "terra incognita" range, between 100 m and 1 km. We used the MYNN 1D PBL parameterization in this study and it resulted in unphysical secondary circulations. In contrast, the simulation with the 3D PBL parameterization correctly maintained homogeneity in one horizontally-homogeneous direction while capturing the dynamical effects of the heterogeneity in the other horizontal direction. We have demonstrated that the results obtained using the 3D PBL parameterization are consistent with the averages from an ensemble LES.

We have carried out high-resolution LES of a flow over the WFIP2 filed study area and compared the results with observations at the BPA towers and by the Vaisala sodars. In general, the LES resulted in lower MAE and RMSE when compared to the mesoscale simulation results. This is consistent with the expectation that a high-resolution LES better captures complex terrain effects on the flow.

Finally, we have assessed the performance of an intermediate form of the 3D PBL parameterization which utilizes the boundary-layer approximation in order to directly solve a system of linear algebraic equations for all the turbulent stresses and fluxes. For that purpose, we have used observations during

the “ten-day” retrospective period from August 13 to August 24, 2016. Horizontal normal turbulent stress components are underpredicted by the 3D PBL as well as one of the shear stress components, $\langle u'v' \rangle$, while the two turbulent shear stress components $\langle u'w' \rangle$ and $\langle v'w' \rangle$ are accurately predicted. Overall, the 3D PBL parameterization results in slightly better correlation with the observation, essentially the same RMSE, and significantly improved variance.

At present the 3D PBL parameterization implemented in WRF is Level 2 according to Mellor and Yamada (1982) classification. This means that the TKE is estimated using a diagnostic equation. However, according to the Level 2 parameterization all the turbulent stresses and fluxes depend directly on the TKE. Solvability of the linear algebraic equations critically depends on the exact level of TKE potentially resulting in numerical instabilities. It can be expected that introducing a prognostic equation for the TKE, or in other word Level 2 ½ parameterization, would alleviate some of the numerical instability issues encountered when the full 3D PBL parameterization is used. Following implementation of the prognostic equation for TKE (or more precisely double TKE) LES results will be used to guide further development of the 3D PBL parameterization including calibration of different model parameters. Finally, the “ten-day” retrospective period mesoscale simulations should be carried out to assess the performance of the full 3D PBL parameterization. Based on the results of the assessment additional parameter tuning may be necessary to obtain optimal performance. Further improvements of the 3D PBL parameterization may include better treatment of the boundary top entrainment.

3.2.2 Analysis of Wind Farm Wake Parameterization

Wakes from individual wind turbines reduce the power available to turbines located downwind. Similarly, wakes from aggregations of wind turbines have been observed to persist many kilometers downwind, reducing the wind energy available to wind plants located downwind and aggravating the challenges of forecasting winds in regions of extensive wind development. Wakes are particularly long-lived in stable nighttime conditions.

Since 2011, WRF has included an open-source wind farm parameterization (WRF-WFP) that represents the aggregate effect of wind turbines on the flow via an elevated source of turbulent kinetic energy and an elevated momentum sink or drag (Fitch et al., 2012; Fitch et al., 2013; Lee and Lundquist, 2017). This parameterization has been loosely validated in comparison to large-eddy simulations, wind tunnel experiments, and with respect to near-surface temperature observations, and its performance was examined for the offshore Horns Rev wind plant (Jiménez et al. 2014).

In WFIP2, the team at CU analyzed the performance of WRF-WFP using data from the field study and explored the sensitivity of the model the implementation of a rotor-equivalent wind speed. See Redferen (2019) for a detailed discussion. The current version of the WFP applies the hub-height wind speed to all layers instead of allowing each vertical layer to have its own wind speed. However, Wagner et al. (2011) proposed the use of a rotor-equivalent wind speed (REWS) instead of hub-height wind speed for power curve evaluation, and Choukulkar et al. (2015) modified the formulation of REWS to include wind direction and turbulent kinetic energy (TKE). An implementation of REWS was introduced in WRF-WFP to replace use of layer-unique wind speeds, and the effect on performance was tested.

Generally, the inclusion of REWS or REWS modified by wind direction and TKE introduces very subtle effects compared to the standard WRF-WFP. In majority of cases tested, wind speed deficit plots were indistinguishable, and the only scenario where a significant difference emerged was during cold-pool mix-out, when with highly non-linear wind shear across the rotor plane, REWS was more accurate.

We focus here on the behavior of WRF-WFP in two easterly flow cases involving the Biglow Canyon wind plant near Wasco. The model terrain used in both is shown in Figure 21. The first is a summertime event from August 16-18, 2016, when a cold frontal surge reached the study area as a surface cyclone built across Montana. The Physics Site experienced easterly/northeasterly winds, and an interesting set of waves developed in the simulation, as shown in Figure 22, which shows the difference in wind speeds

across the model domain between simulations that are identical except for the presence of the WFP. The wave activity is consistent with gravity waves propagating from the wind plant, as proposed by Smith (2009). Strikingly, the magnitude of the wave pattern exceeds that of any more static wake in the lee of the wind plant.

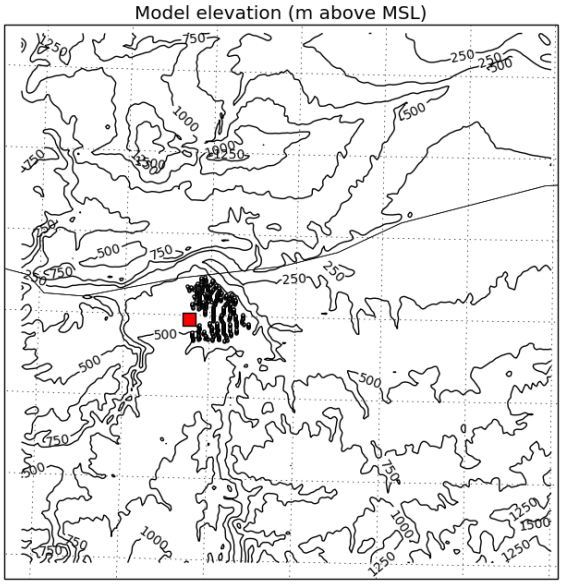

Figure 21. Model terrain elevation for Wasco simulations.

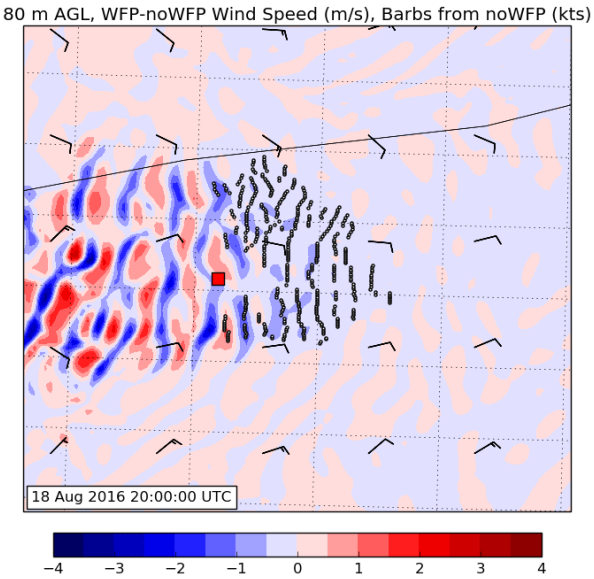

Figure 22. Difference in 80m wind speed at 2016-08-18 20Z between two WRF simulations that are identical except for the presences of the wind farm wake parameterization. Small dots show the presences of wind turbines.

The second case, from Nov 23-25, 2015 features more northerly flow, but again displays considerable wave activity, as shown in Figure 23. In this case, the wakes appear stronger (consistent with the flow aligned with the rows of turbines), but the wakes aren't much stronger than manifestation of wave activity.

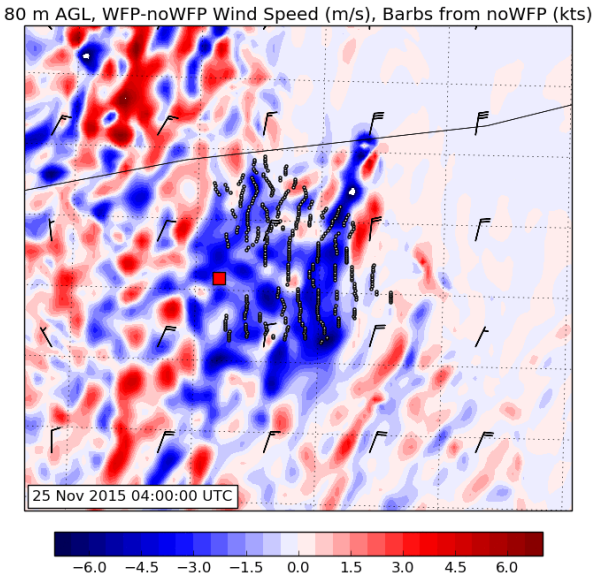

Figure 23. Difference in 80m wind speed at 2015-11-25 04Z between two WRF simulations that are identical except for the presences of the wind farm wake parameterization. Small dots show the presences of wind turbines.

These simulations are not conclusive, but suggest that low-level gravity wave propagation in the vicinity of a wind plant could play a significant role in the variability of generation.

3.3 Observations and Analysis

3.3.1 Observational Processing and Quality Control

The University of Notre Dame group developed quality control (QC) procedures to filter the data and generate secondary products from UND instrumentation. These products are included in the b0 dataset of the DAP. They include:

a) Vertical profiles of horizontal wind speed and wind direction for the Scanning Lidar at Boardman

b) Turbulent momentum and heat fluxes for the PS01, PS02, PS06 and PS11 towers at the Physics Site

c) QC'd data for the ceilometer at Wasco, and the Sodar-RASS and microwave radiometer at Rufus.

For detailed information regarding the QC procedures applied to the UND instrumentation, their range of operation, and metadata please refer to the A2E portal.

3.3.2 Analysis of Physics Site Data

During WFIP2, the Physics Site (hereafter PS) was heavily instrumented to capture complex flow patterns that regularly arise over complex terrain. The topography and location of meteorological towers deployed at PS are shown in the map of Figure 24. The map was generated from the National Elevation Dataset (NED) (https://nationalmap.gov) with a horizontal resolution of 1/3 arc second (≈10 meters).

The site is characterized by a gentle hill, with elevation variability of about 80 m over 4 km in the East-West direction. The slopes vary between 0 and 20 degrees. The horizontal black line on the map denotes the East-West transect along the main slope, where four UND sites (PS01, PS02, PS06 and PS11) were located at different distances from the hill crest, with PS01 being the furthest away (about 3 km) followed by PS02 (≈2.3 km), PS06 (≈0.75 km) and PS11 (≈0.3 km). Exact locations, type of instrumentation, and measurement heights for the sites are listed in Table 9 and

Table 10.

The height of the hill with respect to the terrain elevation at PS01 is about H=70 m. The terrain slope between PS01 and PS02 is only ≈ 3.7^{o}. However, the terrain immediately south of the ridge where PS01 and PS02 were ilocated is quite steep, with slopes up to 20 degrees and a change in elevation of approximately 40 meters over about 0.75 Km in the North-South direction. PS06 and PS11 are located uphill, with PS06 being situated 20 m below the hill top and aligned with the West-East transect. PS11 is 12 m below the hill top and located 30 degrees West of North.

Table 9. UND tower locations

Tower	Latitude (deg)	Longitude (deg)	Tower Height (m)	Elevation ASL (m)
PS01	45.6374	-120.6799	10	428
PS02	45.6383	-120.6716	17	445
PS06	45.6379	-120.6508	21	474
PS11	45.6393	-120.6460	10	484

Table 10. Instruments and measurement heights for UND Towers.

Tower	3D Sonic Anemometer heights (m)	T/RH sensors heights (m)	Gas Analyzer (LiCOR) height (m)
PS-01	3,10	3,10	3
PS-02	3,10,17	3,17	-
PS-06	3,10,21	3,10,21	-
PS-11	3,10	3,10	-

Multiple remote sensing instruments were deployed at the Rufus, Wasco and Boardman sites. The instruments complemented the other suite of instruments deployed at each of these locations by the WFIP2 team.

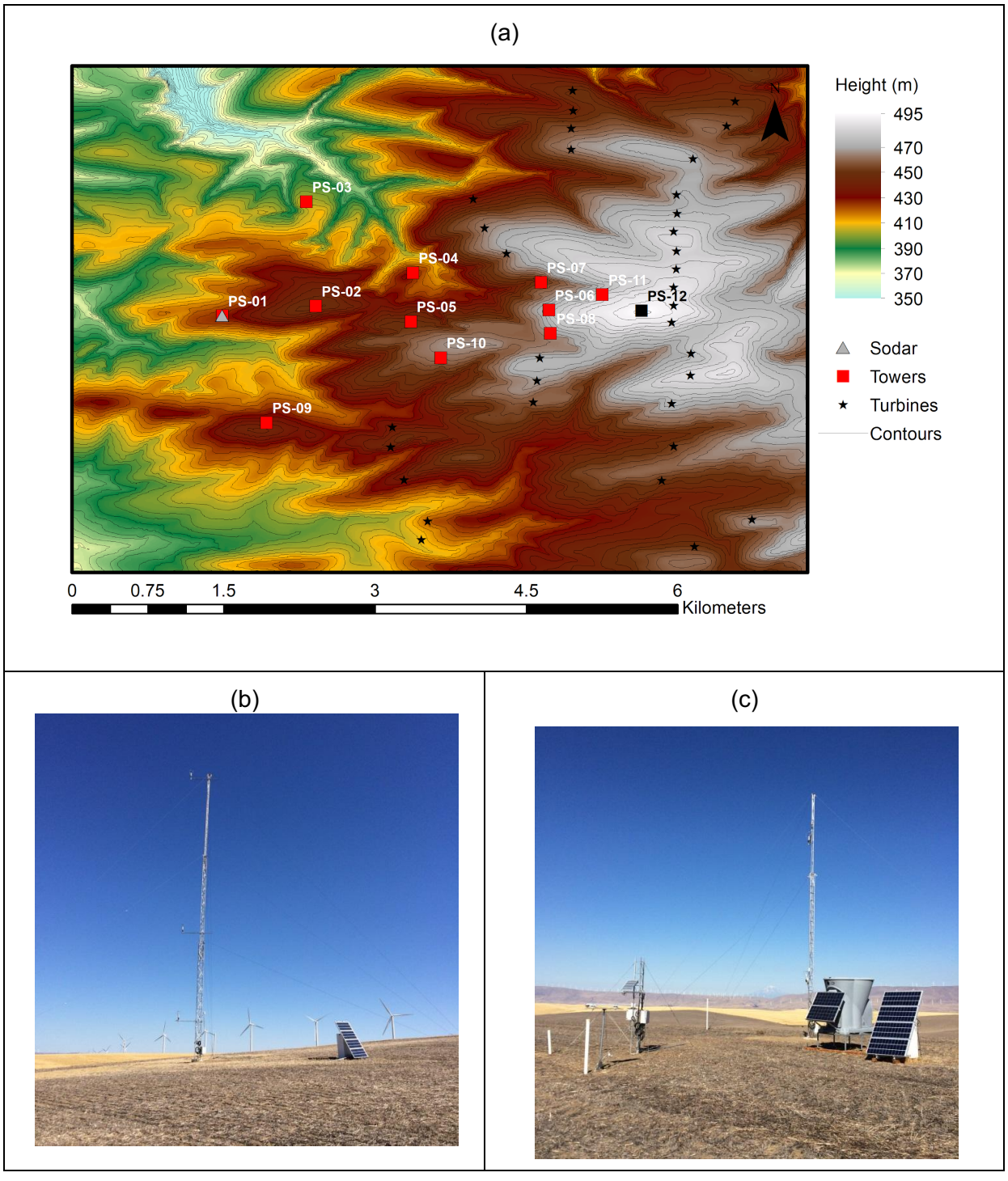

Figure 24. (a) The topography of the Physics Site and the location of towers and wind turbines (b) UND tower PS06 and (c) UND tower PS01 and co-located instrumentation of collaborators (Sodar and solar/soil measurements).

3.3.2.1 Terrain Complexity

In general, landforms are referred to as “complex terrain” based on different combinations of slope and elevation (Stipersky and Rotach 2016), but no widely accepted measure of “terrain complexity” exists, making its definition somehow subjective or indeterminate (Petersen et al. 1998). An objective measure of steepness or ruggedness of the terrain can be classified using the Ruggedness Index (RIX). RIX is defined as the percentage fraction of the terrain that is stepper than a critical slope, usually 0.3 (Wood 1995; Petersen et al. 1998). The index was proposed as a coarse measure of the extent of flow separation, and thereby the extent to which the terrain violates the requirements of linearized flow models, such as WASP (Bowen and Mortensen 1996).

Landscapes may be characterized by the following RIX values: flat and hilly 0%, more complex about 10% or less, and mountainous from about 10 to 50% or more (Petersen et al. 1998; Bowen and Mortensen 1996). The validity of the RIX builds upon the work of Wood (1995), who estimated a critical slope θ_{cr} associated with the onset of mean flow separation as a turbulent boundary layer passes over a hill under neutral stratification conditions.

To calculate the RIX for WFIP2, we used the NED in combination with ArcGIS v10.4. Slopes and RIX calculations were performed for the Inner-Domain of WFIP2, shown in Figure 25, and Physics Site (PS) shown in Figure 26. The physics site is approximately the size of a single high-resolution grid cell for an operational NWP model. The PS was meant to identify physical processes that are enhanced by the terrain and accounts for turbulent fluxes variability at the sub-grid scale of NWP models.

For the Inner-Domain of WFIP2, the slopes have an average of 9º, a standard deviation of 10º and a range that fluctuates between 0 and 80º, as indicated in Figure 25 (a). The major topographic features that are likely to modify the flow on the Inner-Domain of WFIP2 are the Macks Canyon, located south of the Columbia River and about 20 Km west of the PS, characterized by steep walls surrounding the Deschutes River all the way up to its confluence with the Columbia River and Gordon’s Ridge, about 20 Km South-West of the PS. Both of these topographic features may modify the flow on a smaller scale at the PS for strong westerly winds, Other possible features that modify the flow are the Cottonwood Canyon, located south of the Columbia River and about 10 Km East of the PS, and whose steep walls encompasses the John Day River, that might be important in the presence of strong easterly winds and the steep walls of the Columbia River in the presence of strong northerly winds. These landscapes are stepper than 0.3 (~ 17 º), and thus te flow separation is expected to occur therein, as indicated by the bright green-colored areas in Figure 25 (b). The RIX or the percentage of the area steeper than 0.3 for the Inner-Domain of WFIP2 is of 17%, and thus the terrain can be classified as *Mountainous*.

The results on a smaller scale at the PS are shown in Figure 26 (a) and Figure 26 (b). The terrain complexity of the PS is smaller than for the overall Inner-Domain. The RIX for the PS is 2% and since flow separation may still occur, the terrain is classified as *Complex*. The topographic features that account for most of the PS complexity are a) canyon with a longitudinal axis oriented in the NW-SE direction that extends over the upper half of the PS domain (in the presence of north-easterly winds and b) series of smaller canyons with their longitudinal axis oriented in the NE-SW direction that may be important to consider when the winds are northwesterly.

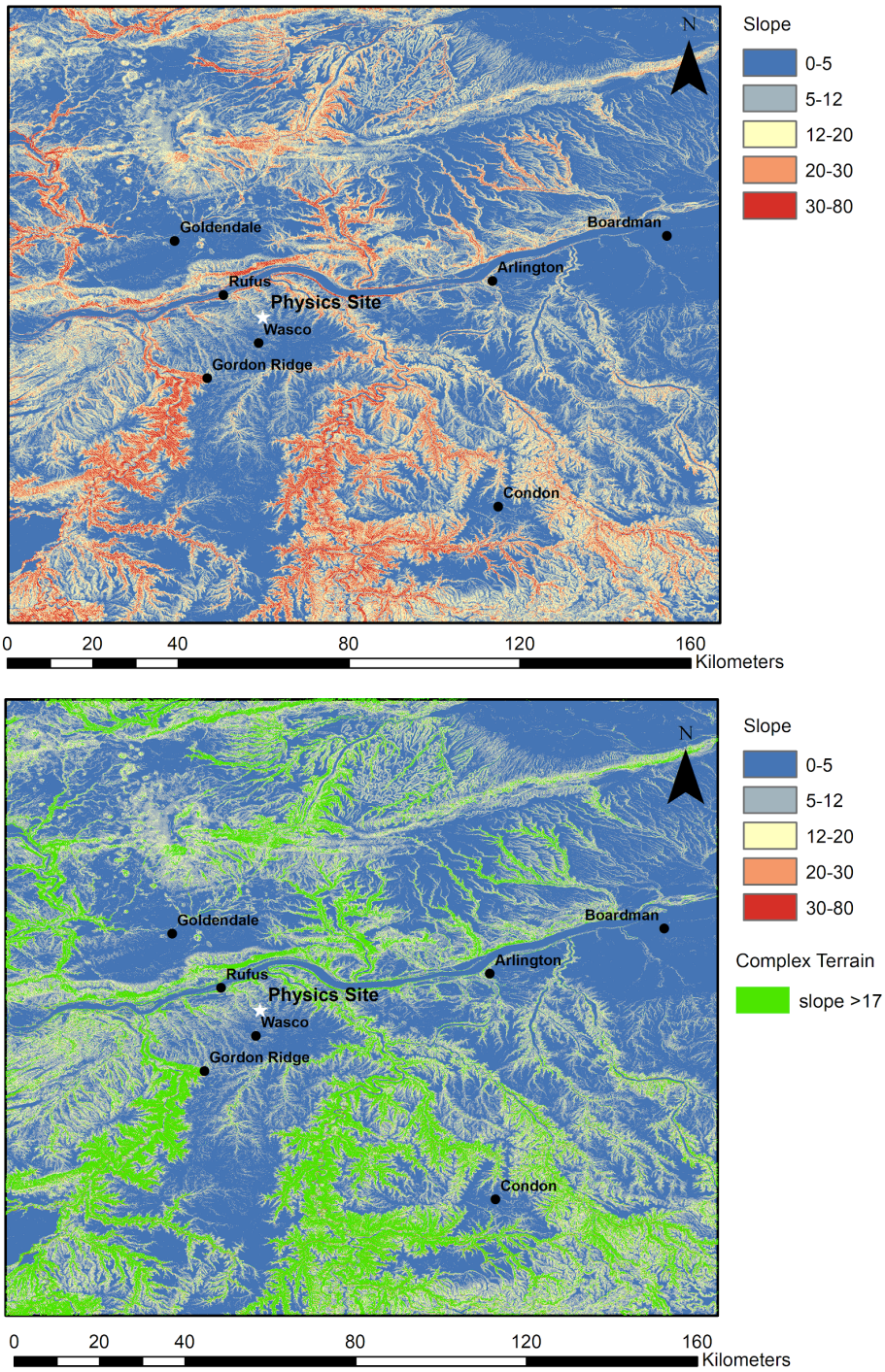

Figure 25. a) Top panel: Map of slopes for WFIP2 Inner Domain, the location of the PS is indicated by the white star. In addition, the main locations where instrumentation was deployed are indicated by the black dots. b) Bottom panel: Same as 1), but with the areas of slopes greater than 17 degrees overlaid to it and indicated by the bright green color. These areas represent possible zones of flow separation that might have an effect on the flow at the Physics Site. The RIX for the entire WFIP2 Inner-Domain region is of 17%, thus the terrain is *Mountainous.*

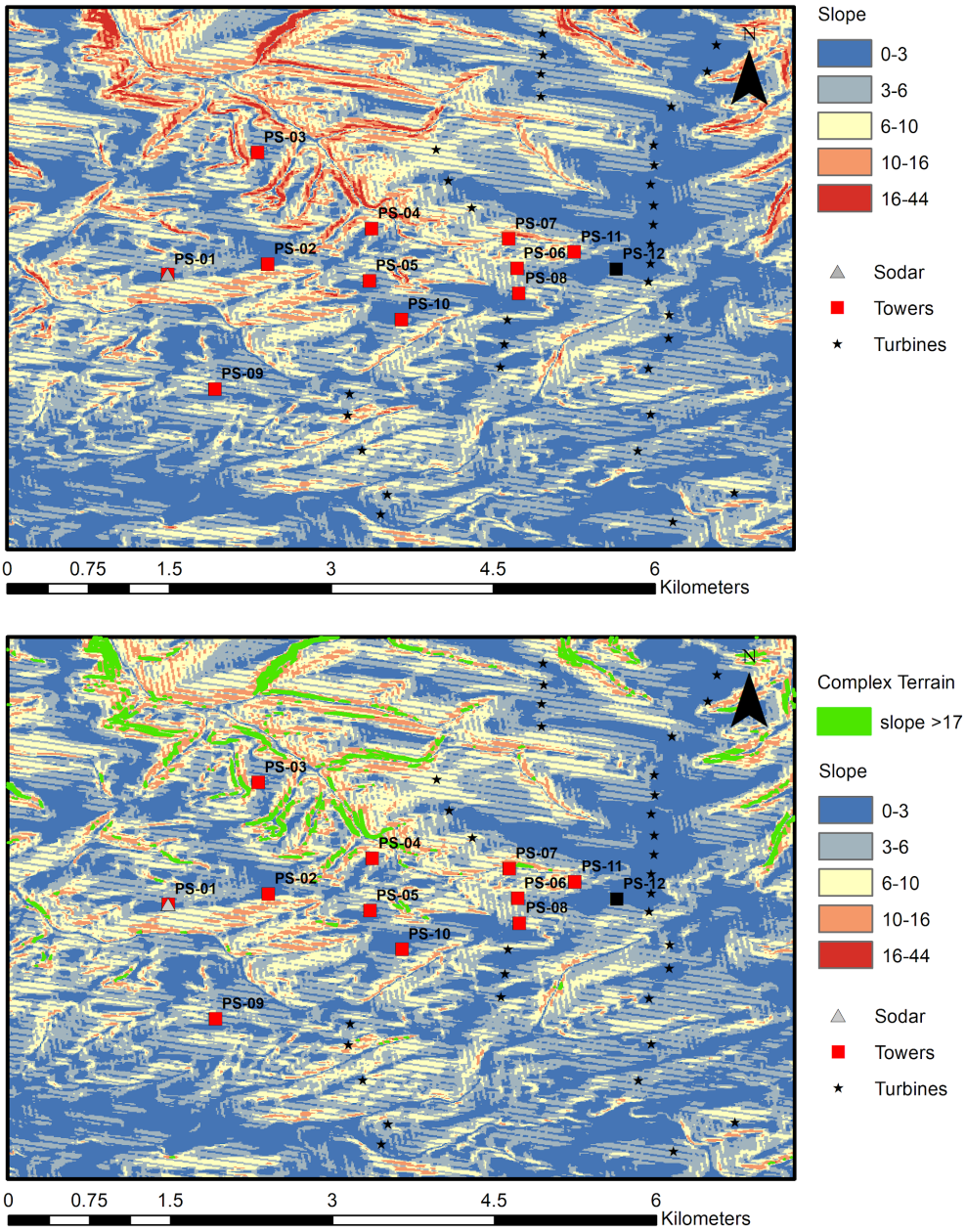

Figure 26. a) <u>Top panel:</u> Map of slopes for the Physics Site area, the tower locations are indicated by the square symbols and the location of wind turbines are indicated by the black-colored stars. A sodar was installed at PS01 to measure wind velocity profiles of the incoming flow into PS, and it is indicated by the gray-colored triangle. b) <u>Bottom panel:</u> Same as 1), but with the areas of slopes greater than 17 degrees overlaid and indicated by bright green. These areas are possible flow separation zones that might affect flow at the Physics Site. The RIX for the Physics Site Domain is of 2%, thus the terrain at the PS is *More Complex.*

Even though a terrain classification based on a RIX indicates some important features of the terrain, RIX will only be of limited value when analyzing the variability of turbulent fluxes at the subgrid scales of NWP models. Note that RIX depicts the terrain “complexity” only in the perspective of the onset of mean flow separation regardless of any other topographically influenced processes that may modify the flow. Therefore, for complete characterization of complex terrain, a composite index that considers an array of processes should be considered, which ought to include, in addition to flow separation, processes such

as turbulence generation mechanisms, buoyancy production, local flow inhomogeneities and non-stationarities.

3.3.2.2 Second order moment analysis at Physics Site

The structure of turbulence over flat homogeneous terrain and under various atmospheric boundary layer (ABL) conditions has been studied extensively (e.g. Kaimal et al. 1972, 1976; Piper and Lundquist 2004), partly motivated by the development of Monin-Obhukov Similarity Theory (MOST), which states that the correct dimensionless wind and temperature gradient in the surface layer, $(\kappa z/u^*)\partial U/\partial z$ and $(z/T^*)\partial T/\partial z$, respectively, are dependent exclusively on the parameters $g/{T_0}^*, u^*, q/(\rho C_p)$and the height z. Therefore, only one dimensionless coefficient $\frac{z}{L}$ can be formed, with L being the Obukhov length, κ the von Karman constant, g the gravity, ${T_0}^*$ the temperature scale, q the kinematic heat flux, C_p the specific heat and ρ the air density.

The structure of turbulence in complex terrain, however, remains relatively unexplored, in spite of its paramount importance for a wide range of applications, i.e. urban air pollution, dispersion in cities, wind energy harvesting, aviation, firefighting and NWP (Wood, 1999; Vecenaj et al., 2010; Fernando et al., 2015). Current spatial and temporal resolutions of NWP models can be as high as 1 km and 10 s, respectively, thus allowing to represent intricate topographical features and to simulate canonical flows that are frequently observed in complex terrain (Skamarock et al., 2005; Papanastasiou, 2010). Nonetheless, the Kolmogorov scales in the atmosphere are of ~1mm and ~1s, respectively (Fernando, 2013), and there is a whole host of scales that remain unresolved by the NWPs. This is usually accomplished using standard surface-layer MOST, regardless the characteristics of underlying terrain (Garratt, 1992).

To study the performance of MOST in complex terrain surrounding PS, June-July-August (JJA) 2016 was selected as a case study. The 15 min averaged turbulent fluxes calculated from the UND and PNNL towers were used for analysis. The spatially averaged non-dimensional root mean square velocities normalized by the friction velocity; σ_u/u^*, σ_v/u^*, σ_w/u^* were plotted as a function of the non-dimensional height z/L for data taken at heights 3 m and 10 m, and the results are shown in Figure 27 and Figure 28, respectively. Positive values of z/L corresponds to the stable stratification whereas negative values correspond to unstable stratification. The spatially averaged non-dimensional RMS velocity components for the PS are indicated by the red dots in the figures. The solid black line in the figures shows canonical similarity functions derived for flat and homogeneous terrain (FHT); see Table 11. The blue line in the figures correspond to the best fit for the measured data, parameterized as:

$$\sigma_i/u^* = \alpha_i(1 + \beta_i |\xi|)^{1/3}$$

The previous parameterization was adopted from de Franceschi et al. (2009), whihch was proven to be successful in parameterizing the non-dimensional root mean square velocities over an Alpine Valley (de Franceschi et al. 2009). Moreover, it is in general agreement with MOST parameterizations such as the ones developed by Panofsky and Dutton (1984). The values of the constant $\alpha_i$and β_i obtained for the PS data during the period JJA are given in Table 12 for measurements at 3m and in Table 13 for measurements at 10m.

During the summer, the wind climatology is predominantly from the West. In this analysis, the data was filtered to include wind directions between 221 to 316 degrees. This wind direction sector is undisturbed by the wake effects of wind turbine locate near the PS. From Figure 27 and Figure 28 it can be observed that σ_u/u^* and σ_v/u^* are significantly higher than that predicted for Flat and Homogeneous Terrain (FHT) during unstable conditions. This trend is observed at 3m and 10m. At both heights the data is quite scattered and a relationship between σ_u/u^* and σ_v/u^* exclusively based on the stability parameter z/L doesn't seem to hold for the PS area. Interestingly, σ_u/u^* and σ_v/u^* are generally overpredicted during stable conditions at both 3m and 10m, and thus the buoyancy effects that reduce the TKE is not captured very accurately by MOST over Complex Terrain. Finally, the values of σ_w/u^*measured at the

Physics Site are smaller than FHT predictions during unstable conditions, but somewhat larger during stable conditions. Nonetheless, the deviations from MOST for σ_w/u^*were smaller compared to σ_u/u^* and σ_v/u^*.

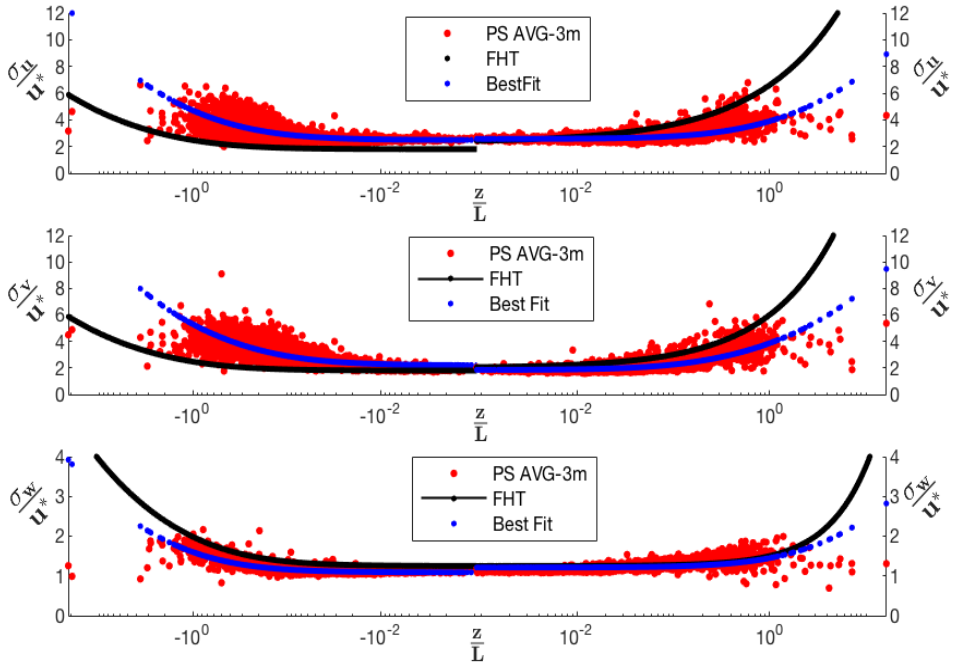

Figure 27. Physics Site - spatially averaged standard deviations for the wind velocity components normalized by the friction velocity, σ_u/u^*,σ_v/u^*,σ_w/u^*, plotted against the non-dimensional similarity variable z/L^ for the period June, July, August 2016 at 3m.

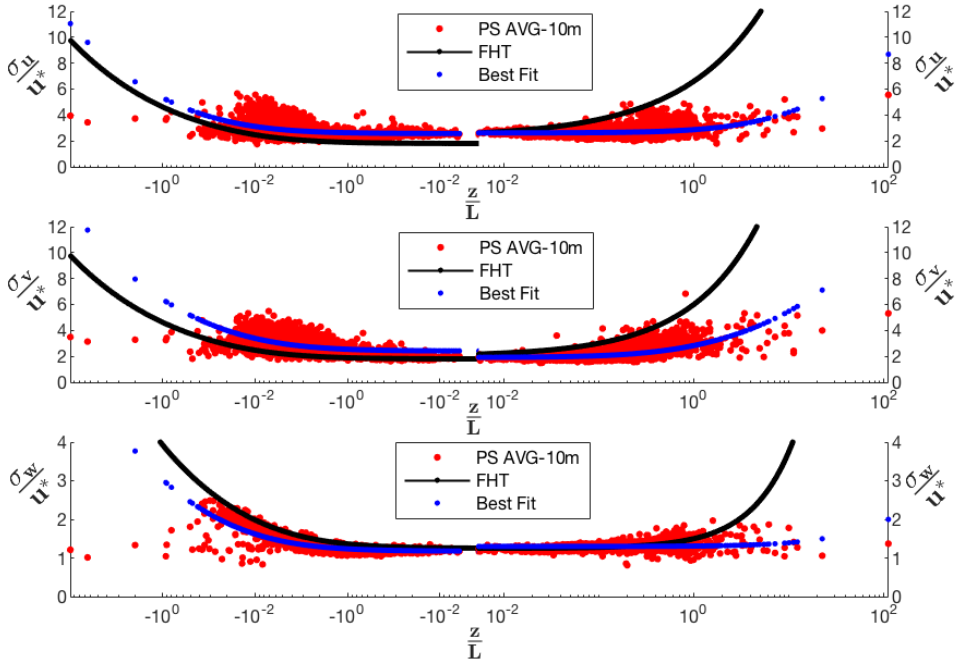

Figure 28. Physics Site - spatially averaged standard deviations for the wind velocity components normalized by the friction velocity, σ_u/u^*,σ_v/u^*,σ_w/u^*, plotted against the non-dimensional similarity variable z/L^ for the period June, July, August 2016 at 10m.

Table 11. Similarity Functions for Flat and Homogeneous Terrain (FHT).

Variable	Stability Parameter, $\xi = z/L$	Expression	Reference
σ_u/u^*	$\xi > 0$	$1.8(1+1.6\xi)^{1/3}$	Zhang et al 2001
	$\xi < 0$	$2.3 + 4.3\xi^{1/2}$	Pahlow et al 2001
σ_v/u^*	$\xi > 0$	$1.8(1+1.6\xi)^{1/3}$	Zhang et al 2001
	$\xi < 0$	$2.4\xi^{0.6}$	Pahlow et al 2001
σ_w/u^*	$\xi > 0$	$1.25(1+0.2\xi)$	Kaimal and Finnigan 1994
	$\xi < 0$	$1.25(1+3\xi)^{1/3}$	Kaimal and Finnigan 1994

Table 12. Values of α_i and $\beta_i$obtained for the period JJA 2016 for the PS towers at 3m.

Variable	Stability Parameter, $\xi = z/L$	α_i	β_i
σ_u/u^*	$\xi > 0$	2.57	2.49
	$\xi < 0$	2.49	5.74
σ_v/u^*	$\xi > 0$	1.85	8.10
	$\xi < 0$	2.20	12.86
σ_w/u^*	$\xi > 0$	1.21	0.71
	$\xi < 0$	1.09	2.16

Table 13. Values of α_i and $\beta_i$obtained for the period JJA 2016 for the PS towers at 10m.

Variable	Stability Parameter, $\xi = z/L$	α_i	β_i
σ_u/u^*	$\xi > 0$	2.62	0.32
	$\xi < 0$	2.57	0.80
σ_v/u^*	$\xi > 0$	1.91	8.24
	$\xi < 0$	2.41	1.77
σ_w/u^*	$\xi > 0$	1.30	0.02
	$\xi < 0$	1.18	1.61

3.3.2.3 Turbulence Retrievals

i. Eddy Dissipation Rate (EDR)

Turbulence dissipation rate is a crucial quantity to improve sub-grid scale parameterizations in the MYNN boundary layer scheme (Yang et al. 2017). Doppler lidars provide information about wind field spatial statistics, while providing estimates of turbulence Eddy Dissipation Rate (EDR, ε); see the reviews by Banakh and Smalikho (1997), Frehlich et al. (1998, 2006) and Krishnamurthy et al. (2011). These estimates can be made using temporal data (using temporal spectra), spatial data (using spatial spectra) or a combination thereof (using the estimates of structure function). All EDR estimation algorithms rely on spectral representations of turbulence. Two forms of spectra are equated, as shown in Figure 29; von Karman spectrum at low frequencies and Kolmogorov inertial subrange spectrum at higher frequencies. There are no set values for upper and lower thresholds of the inertial subrange, as they vary with the nature of atmospheric turbulence under different conditions (e.g., stable, unstable and neutral stratification, altitude etc.).

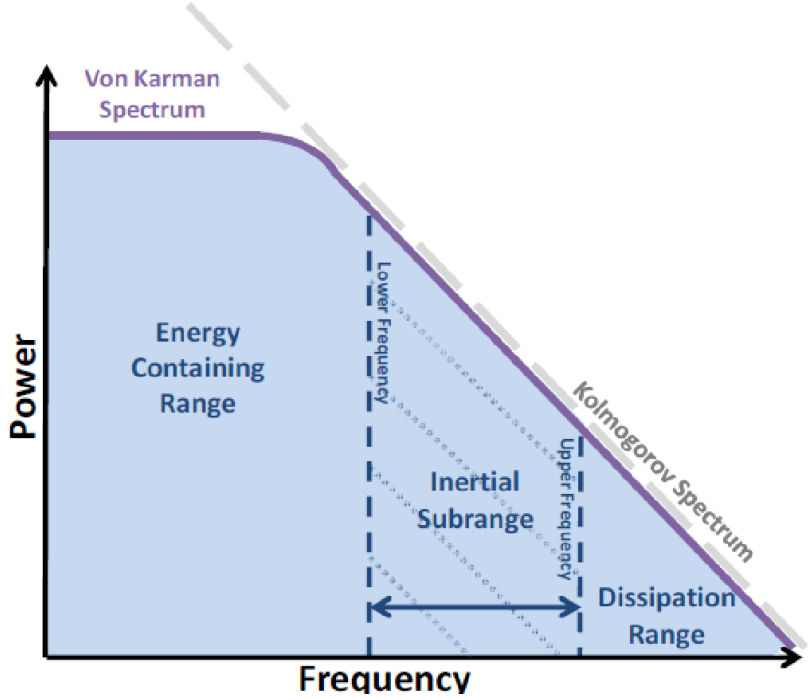

Figure 29. Power spectral representation of turbulence in a log-log plot [Monin & Yaglom (1975)]

The velocity spectrum is the Fourier transform of two-point velocity correlation function, which, in the inertial subrange takes the classical form [Frehlich et al., 2006]

$$E(k) = C\,\frac{2}{3}k^{-\frac{5}{3}},$$

where C is the Kolmogorov Constant. There is a direct relationship between the energy spectrum E(k) and the second order velocity (v) structure function D_v,

$$D_v(s) = \frac{1}{H}\sum_{n=1}^{H} (v(r) - v(r+s))^2,$$

where *H* is the number of samples used for averaging and *r* is the spatial separation. The structure function can be estimated either along the Lidar axis (longitudinal structure function D_{LL}) or along a transverse direction (transverse structure function D_{RR}). They become,

$D_{LL}(r) = C_2(\varepsilon\, r)^{2/3}$,

and

$D_{RR}(r) = C_2'(\varepsilon\, r)^{2/3}$,

C_2' = 4/3 C_2 and C_2 = 2, where the latter is based on previous experimental work (Doviak & Zrinc 1993; also see Chan and Shao 2007). The EDR is then obtained by fitting either the -5/3 slope to the spectrum or the 2/3 slope for the structure function. The units of ε are m^2s^{-3}.

In the analysis below, the transverse structure function is used to estimate EDR, as it deems more reliable based on previous work [Frehlich et al. 2006; Krishnamurthy et al., 2011]. Figure 30 shows the azimuth structure function and Figure 31 shows the vertical profiles of various retrieved turbulence parameters for 21 January, 2017 at 10 hrs. UTC.

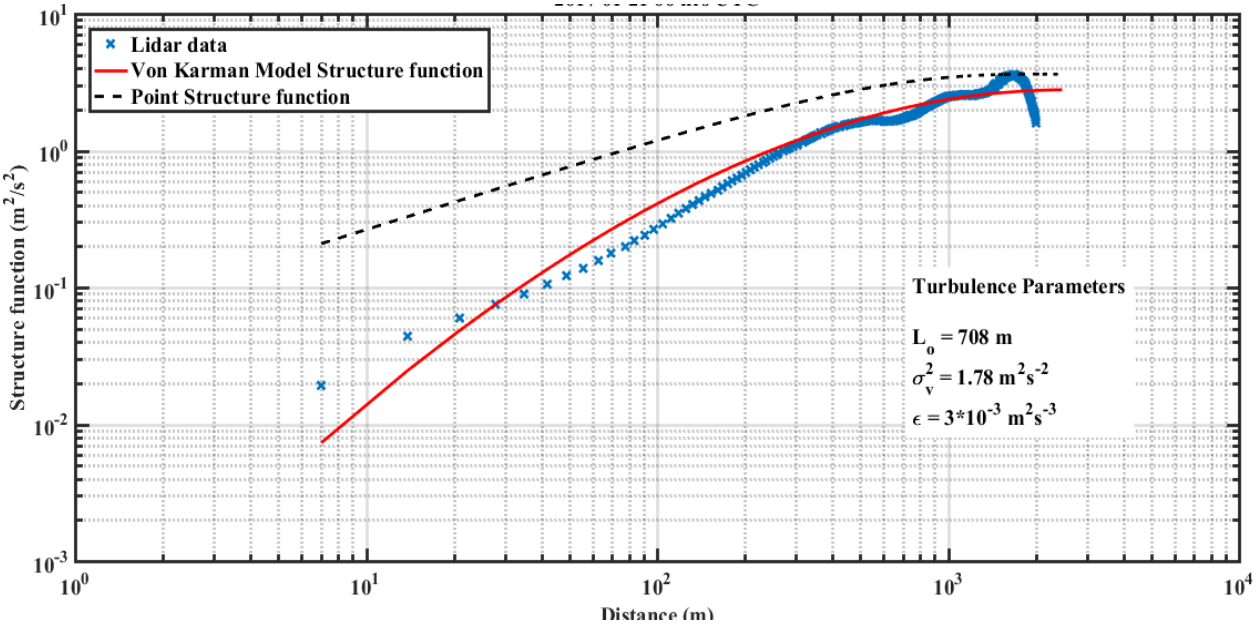

Figure 30. One-hour averaged azimuthal structure function for the Lidar data of January 21, 2017 at 10 hrs. UTC.

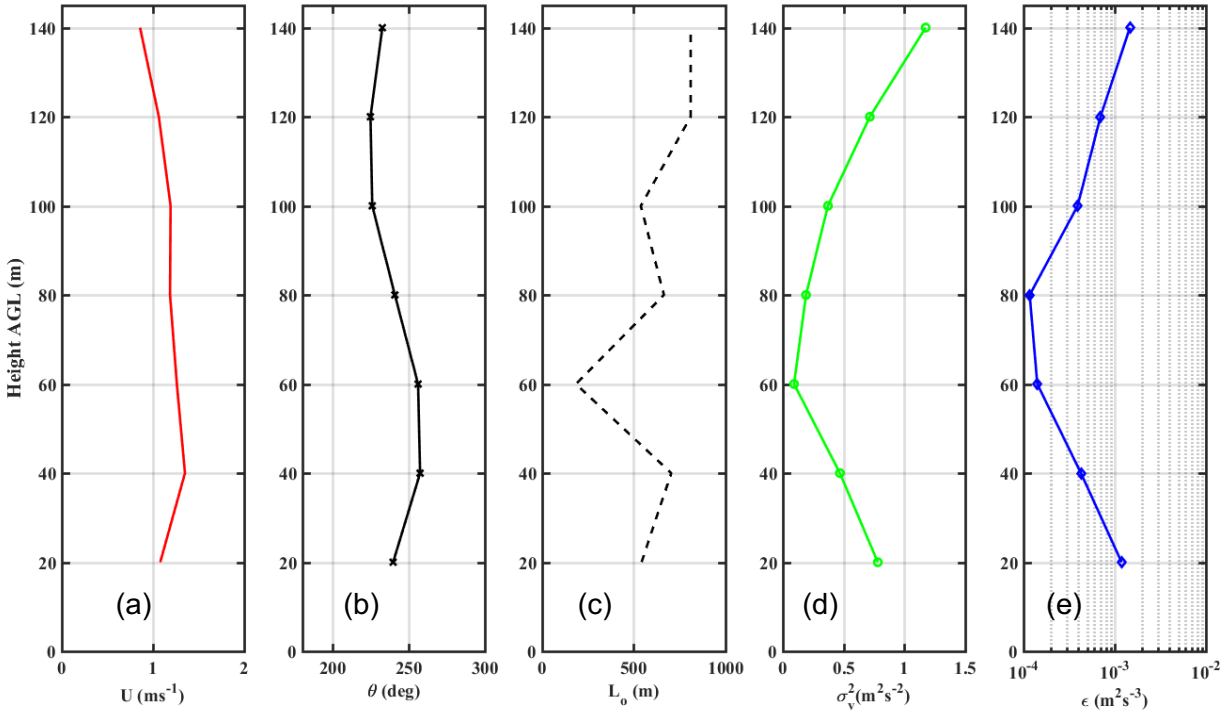

Figure 31. One-hour averaged vertical Profiles of turbulence parameters for January 21, 2017 at 1000 hrs. UTC. a) Horizontal wind speed in m/s, b) wind direction in degrees, c) integral length scale (m), d) velocity variance (m^2/s^2), and e) eddy dissipation rate (m^2s^{-3}).

On this day, a frontal passage approaching from the west developed a high turbulence layer in the bottom 100 m of the atmosphere. Figure 32 shows the radial velocity plots from the scanning Doppler Lidar at different time periods. The wind direction shifts from east to west from 1000 hrs. UTC to 1900 hrs. UTC. Figure 33 shows one hour averaged time series of horizontal wind speed and EDR from 1000 hrs. to 2300 hrs. UTC. During the transition period, a significant increase (up to 10 times) in EDR was observed from 1200 hrs. to 1500 hrs. The winds increased from 2 m/s to 4 m/s, which is the cut-in wind speed for modern day wind turbines. The effect of high wind veer and turbulence during such time periods would result in higher loads on turbine blades as well as variable wake conditions.

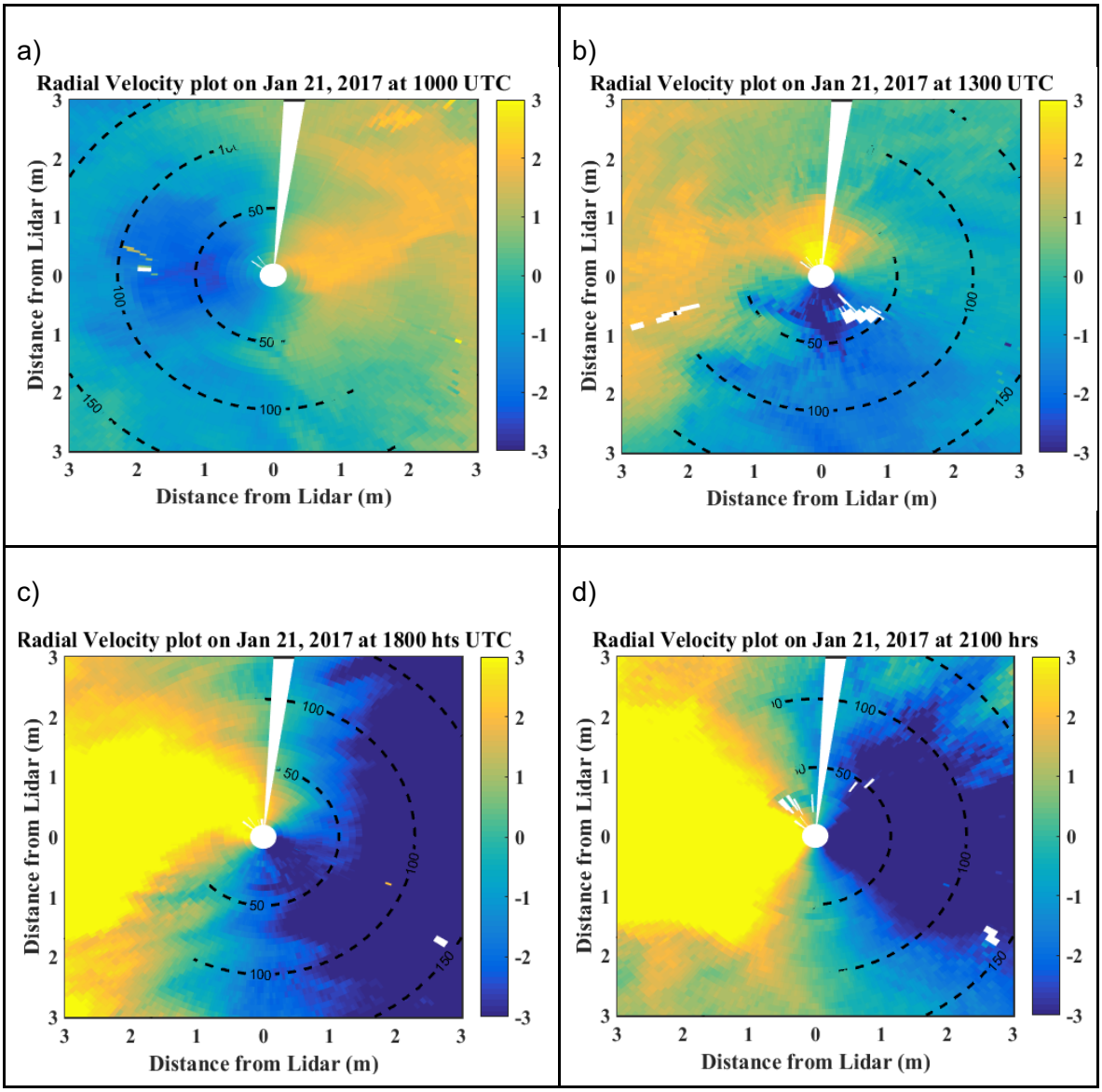

Figure 32. Lidar Radial velocity snapshots at a) 1000 hrs. UTC, b) 1300 hrs. UTC, c) 1800 hrs UTC and d) 2100 hrs. UTC on January 21, 2017, showing the wind veering event. The contour dotted lines show the height above ground level. The radial velocity color convention is yellow wind towards the Lidar and blue wind going away from the Lidar.

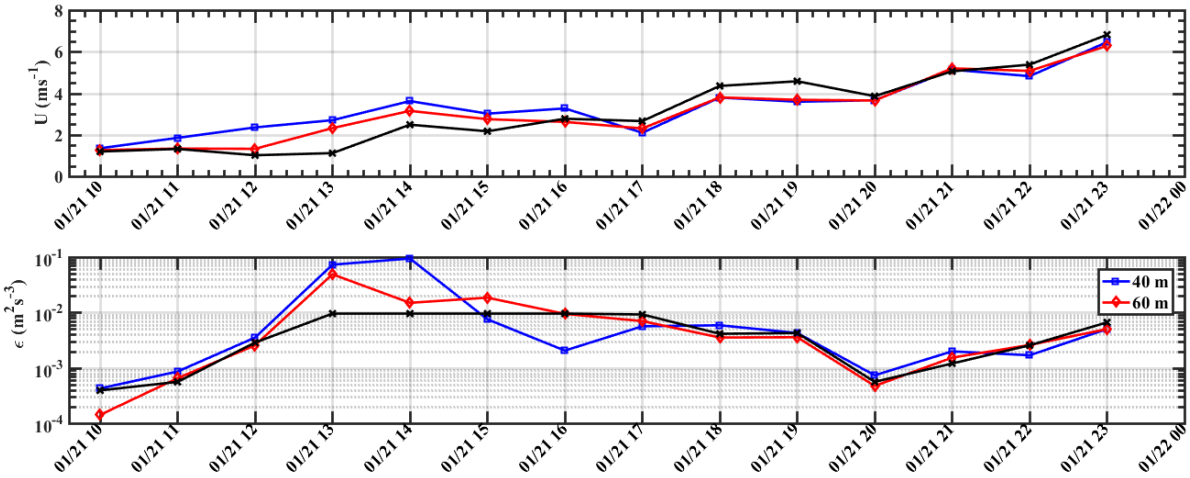

Figure 33. Time series of one-hour averaged horizontal wind speed and eddy dissipation rate at 40m (blue line with open squares), 60m (red line with diamond square) and 100 m AGL (black line with x).

An inter-comparison study of estimating eddy dissipation rate from various remote sensing devices was performed between NOAA, CU Boulder and University of Notre Dame. A profiling lidar, scanning lidar, and wind profiling radar at Wasco, OR were used to measure this quantity, and roughly compared to a nearby (5 km away and 40 m lower in elevation) sonic anemometer on an 80-m tower. More details on the different algorithms shown in Figure 34 can be found in Krishnamurthy et al. (2011), McCaffrey et al. (2017), Bodini et al. (2018), and Wilzack et al. (2019).

The scanning Doppler lidar at Wasco operating in two scan modes were used for turbulence dissipation rates estimates using an a) azimuth structure function method on the 6-degree elevation angle planned position indicator (PPI) scans and b) a vertical velocity spectral slope method on the 3-min vertical stares. The azimuth structure function method estimates of dissipation rate are calculated every 20 m from 30 to 380 m AGL, with decreasing reliability above 250 m AGL, due to limited accuracy of scanning Lidar radial velocity estimates at distant range-gates (since, SNR $\propto$ 1/Range). The spectral slope estimates of dissipation rates are estimated every 100 m from 200 to 10,100 m AGL, with decreasing measurement availability above 1000 m AGL.

Figure 34 shows time series of dissipation rates from the 4 methods and the sonic anemometer, at three heights of interest: a) 80 m AGL where the sonic (roughly) shows the diurnal cycle well-matched by the structure function method of the scanning lidar, b) at 200m AGL where all 4 profiling methods overlap, the structure function method compares well with profiling lidar during daytime convective conditions, and the WPR during nighttime stable periods, showing agreement between multiple instruments at differing parts of the day, and c) at 300 m AGL, where the WPR and spectral slope method of the scanning Lidar match well, as the structure function method of the scanning Lidar is less reliable due to the limitation of accurate scanning Lidar radial velocity estimate at further range-gates (i.e., SNR $\propto$ 1/Range).

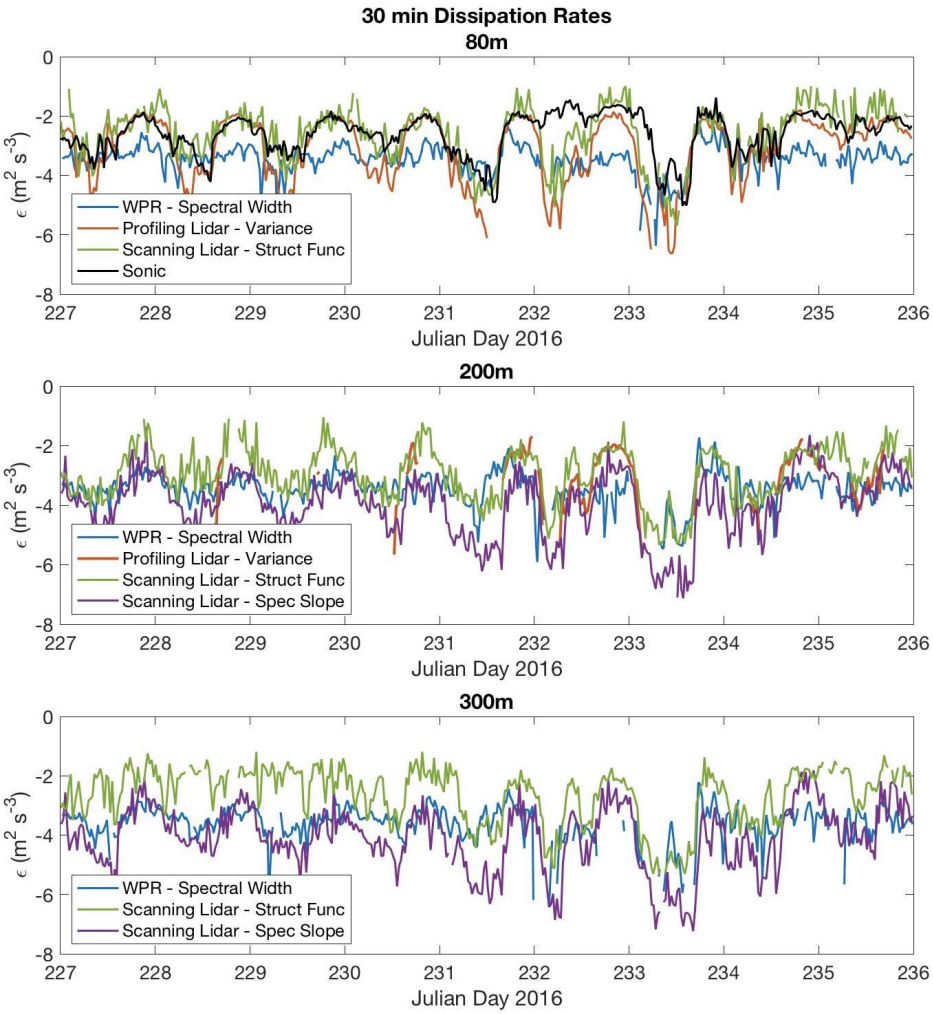

Figure 34. Turbulence dissipation rates at Wasco, OR from the wind profiling radar (blue), profiling Lidar (red) and scanning Lidar using the structure function (green) and spectral slope (purple) methods at 3 heights: a) 80 m AGL, compared to the nearby sonic anemometer (black); b) 200 m AGL, where all four instruments observe; and c) 300 m AGL.

ii. Accuracy of Lidar Retrieval Algorithms

It is important to assess the accuracy of velocity and turbulence retrieval algorithms from scanning Doppler Lidar to reduce the uncertainty and inform the potential users of limitations in using data. As a first step, two velocity retrieval methods based on least-squares were assessed using data from the scanning Lidar located at Wasco airport. Both algorithms use least-squares approach over the 360-degree planned position indicator (PPI) sector to estimate the wind speed and direction. The NOAA algorithm is based on averaging the radial velocity measurements over a 15-minute interval from all scans over the time period and binning them as function of height. The current UND method estimates the wind speed and direction for every scan and then average as a function of height (Krishnamurthy et al. 2013). The two algorithms provide similar estimates, but with a mean wind speed difference averaged over one month of approximately 0.35 – 0.45 m/s, with an averaged wind direction difference of 10 to 12 degrees. The results are shown in Figure 35 and Figure 36.

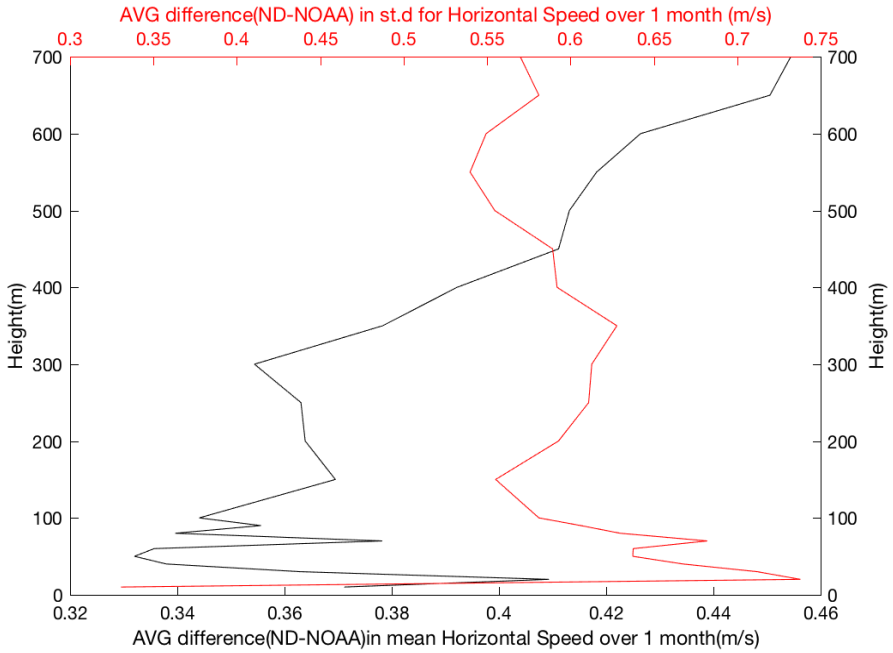

Figure 35. Vertical profiles of the average (black) and standard deviation of horizontal wind speed difference between NOAA and UND algorithms.

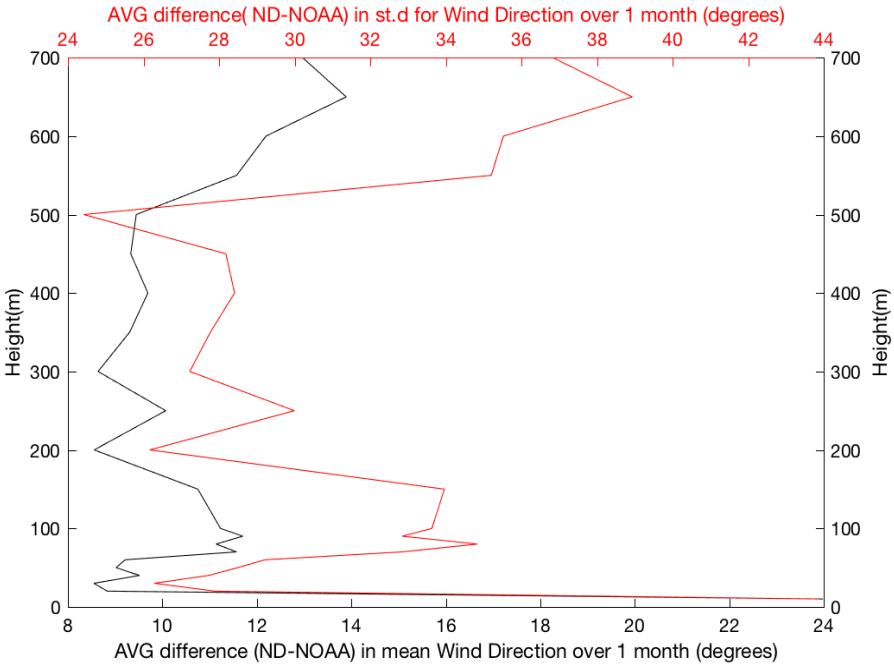

Figure 36. Vertical profiles of the average (black) and standard deviation of horizontal wind direction difference between NOAA and UND algorithms.

Further investigations are in progress by assessing the effect of volume averaging and turbulence on Lidar retrieved products.

3.3.2.4 Case-Studies

The preliminary studies presented here encompass several flows scenarios and are mainly based on the dataset provided by the four UND towers and the scanning Doppler Lidar data collected at Wasco & Boardman sites.

i. Case Study I — Frontal Passage on March 22-April 2, 2016

During the second half of March 2016, the Columbia Gorge area was under the influence of a westerly cold frontal passage. Tower measurement at PS detected the arrival of the front around sunset on March 22 (Julian Day, JD, 82), which was associated with a significant wind power up-ramping. Figure 37 shows 12- day time series (March 21 – April 02) of wind direction, wind speed, air temperature (Tair) and relative humidity (RH) measured at 3m, 10m and 17m at PS02. Raw data from sonic anemometers were processed to remove spikes and tilting effects by applying a maximum velocity filter and the traditional three rotation method (McMillen, 1988) was used to obtain the mean flow in the streamwise direction. Time averages over 15 min were used to calculate turbulence and mean properties of the flow. The figure clearly illustrates the arrival of and penetration down to the surface of the front, as indicated by a swift shift of wind direction from easterly to 260 degrees and the rapid rise of wind speed to ~17 m/s. This westerly high wind regime persisted for about five consecutive days (period A in Figure 37), and after weakening on JD 87, a second event followed that lasted for about 3 days (period B in Figure 37). This regime was finally dissolved on March 30 (JD 90) with the establishment of remarkably calm, light winds and warmer and drier conditions (period C in Figure 37) associated with weak synoptic forcing over the area.

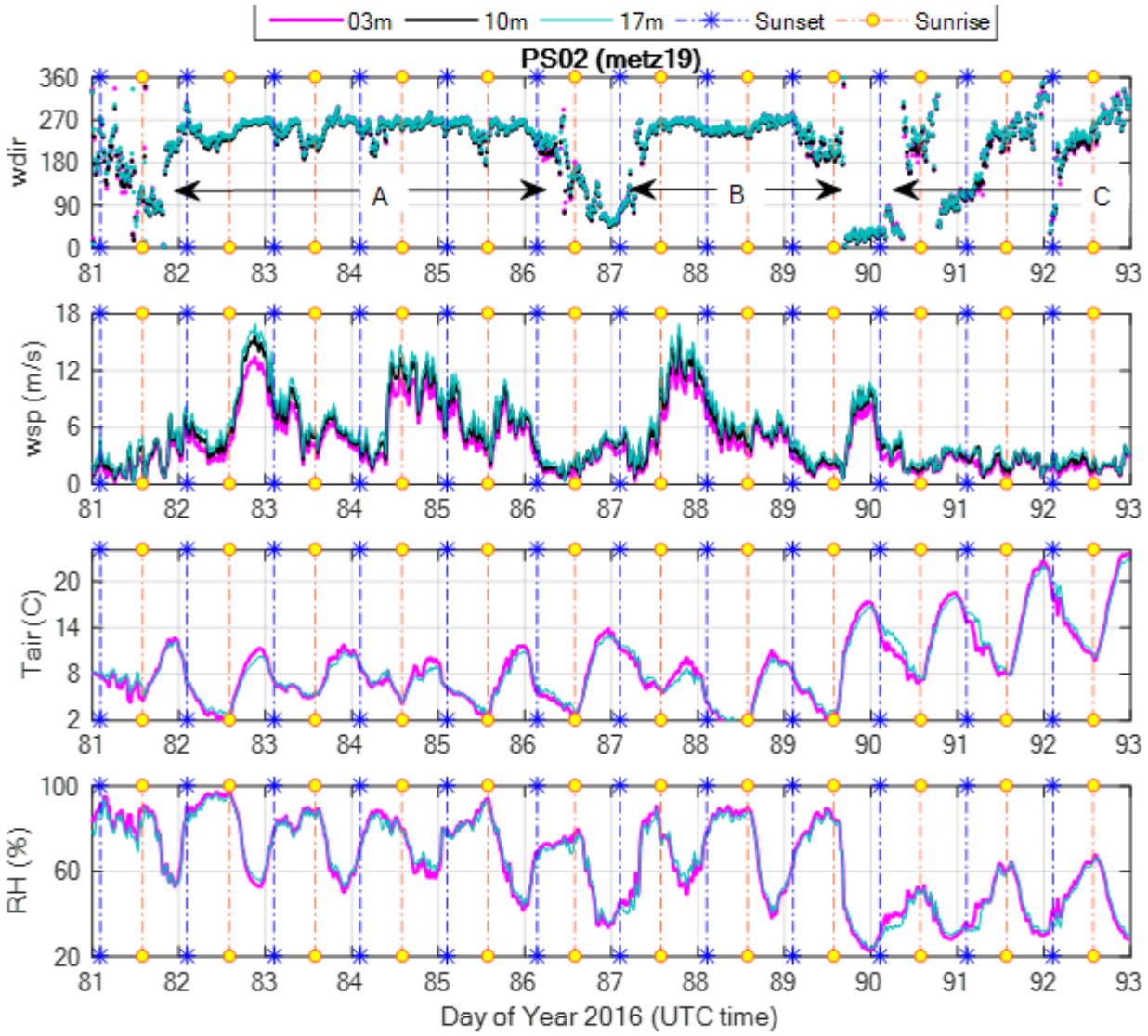

Figure 37. Time series of 15 min averaged wind direction (wdir), wind speed (wsp), air temperature (Tair) and relative humidity (RH) measured at 3m,10m and 17m at PS02 tower.

The analysis of normalized turbulent velocities (Figure 38) for Periods A and C indicate the presence of two disparate turbulence regimes. During Period A, turbulent velocities are low and approximately constant for 5 consecutive days, with both streamwise and spanwise components $\sigma(u) \approx \sigma(v) \approx 0.15U$ and a vertical component $\sigma(w) \approx 0.05U$ (U is the mean speed). On the other hand, turbulent velocities in Period C follow the diurnal cycle, reaching values up to $\approx 0.8U$ and $0.4U$, for horizontal and vertical components, respectively, during the convection. Nighttime turbulence values during period C are comparable to A, although a less stationary trend can be noticed with some remarkable turbulent bursts, which will be subject of further investigations.

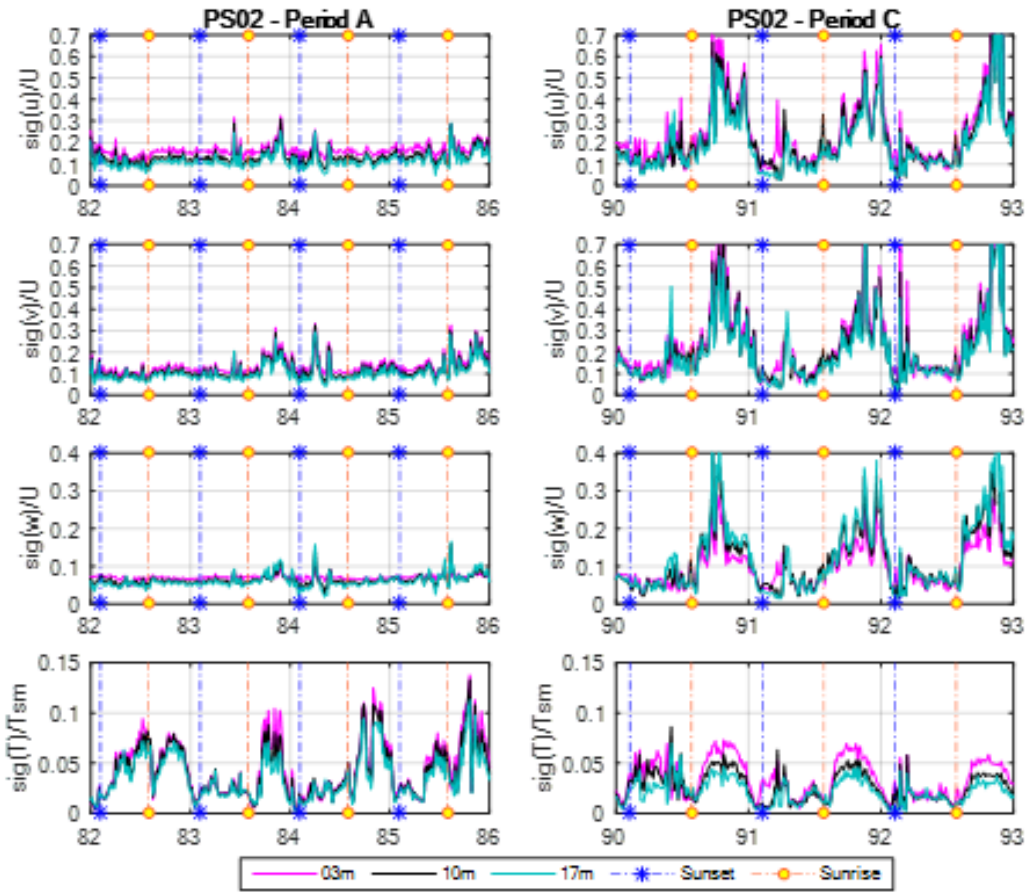

Figure 38. 15-min averaged RMS values of three velocity components (streamwise sig(u), spanwise sig(v) and vertical sig(v)) at PS02 at 3m, 10m and 21m, normalized by the mean speed U at the same height. Normalized RMS values of sonic temperature sig(T)/Tsm are also provided in the bottom panels. The left and right panels refer to Period A and Period C respectively.

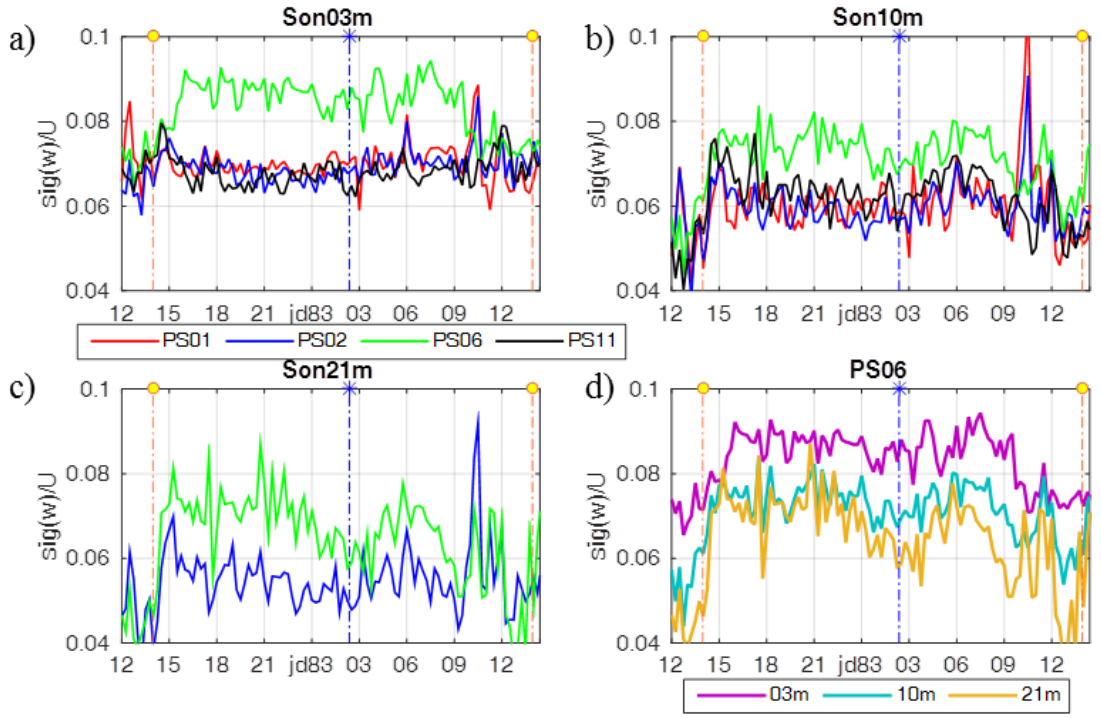

Figure 39. Comparison of normalized RMS vertical velocity at 4 different locations along the East-West transects (PS01, PS02, PS06 and PS11) at heights (a) 3m, (b) 10m, and (c) 20m for a representative 24-hr. segment in Period A. Panel (d) compares three measurement heights of PS06.

Similar analyses were carried out for other UND tower locations along the East-West transect. Results indicate similar trends in terms of both mean and turbulent variables. However, comparison of normalized turbulence velocities at different locations (Figure 39 above) highlights the larger values of normalized vertical turbulent velocities found at PS06 (20 m below the hill's crest) in comparison to other PS locations. In terms of dimensional values, the variation is significant at the lowest measurement level where σ(w) at PS06 is 30% larger than the value measured upstream (PS01). Similarly, σ(u) and σ(v) are 15% larger than the corresponding upstream values. Table 14 and Figure 40 summarize these results based on Period A and also highlight similar trends for PS11.

Table 14. Calculated variation of RMS values of three velocity components (u, v, w) at 3m height for Period A, where 0 denotes the value measured upstream

Variable	PS02	PS06	PS11
$\frac{[\sigma(u)-\sigma(u_0)]}{\sigma(u_0)}$	0%	13%	13%
$\frac{[\sigma(v)-\sigma(v_0)]}{\sigma(v_0)}$	0%	15%	16%
$\frac{[\sigma(w)-\sigma(w_0)]}{\sigma(w_0)}$	1%	29%	22%

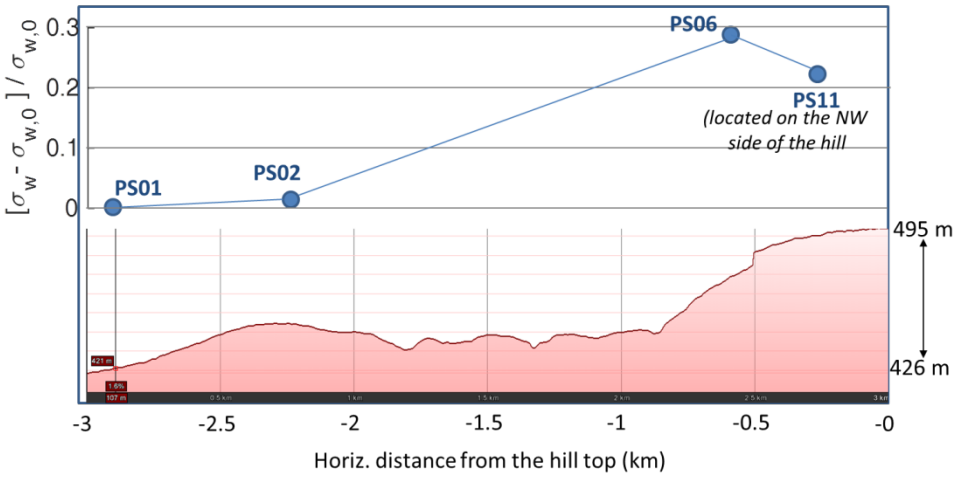

Figure 40. Horizontal cross-section showing calculated fractional σ(w) changes at 3m height for Period A. PS towers positions and terrain shape are also illustrated.

The extension of the analysis to a 3 months-period (February 16, 2016 - April 10, 2016) suggests that the larger values σ(w)/U found at PS06 are always associated to sustained westerly winds, as illustrated in Figure 41. These findings are consistent with the observations of Bradley (1980) and Zeman and Jenssen (1987) as well as in accordance with the rapid distortion theory that predicts an increase in the vertical component of turbulence as the flow approaches the hill crest. On the other hand, consideration of atmospheric stability and Froude number (Fr) for the same dataset indicate that westerly winds would almost always correspond to Fr >1, implying that the flow at the topographic height at PS06 would always flow over the hilltop. Our future work will further investigate these aspects, especially by including data collected at the top of the hill (not available for the first part of WFIP2 field campaign) and considering flow scenarios expected for different atmospheric stability and flow configurations.

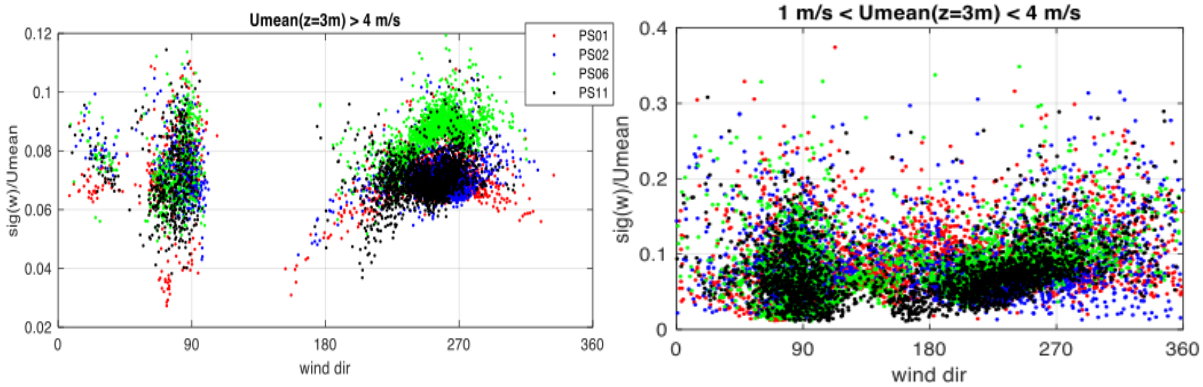

Figure 41. Normalized RMS vertical velocity at 3m for (a) high and (b) low wind speeds as a function of the wind direction. Dataset from February 16, 2016 to April 10, 2016.

ii. Case Study II—Convergent Flows on September 29, 2016

Convergent flows were observed by Scanning Doppler Lidar measurements at the WASCO site (proximity to the Physics site, southeast) on September 29, 2016 at 0500 hrs. Figure 42 shows 15-minute averaged wind and direction profiles from the scanning Doppler Lidar. Winds from northwest (pink color) and from southeast (green) show interesting counter-flowing sheared flow behaviors during the local night time until about noon. It appears that the shear layer is lofted and destroyed with the initiation of strong convection, which allows enhanced momentum diffusion upward. During the counter flow period, we expect strong turbulence at about 1000 MSL, given the shear layer therein with a small mean velocity. We are continuing with studies on the evolution of turbulence in the shear layer as well as the growth of shear layer. Some of the parameters of interest are the turbulent kinetic energy dissipation rate, integral length scale of turbulence as well the RMS velocities. We are applying in-house built algorithms to extract these quantities. Also comparisons are attempted with other instrumentation that is located close to Wasco Lidar.

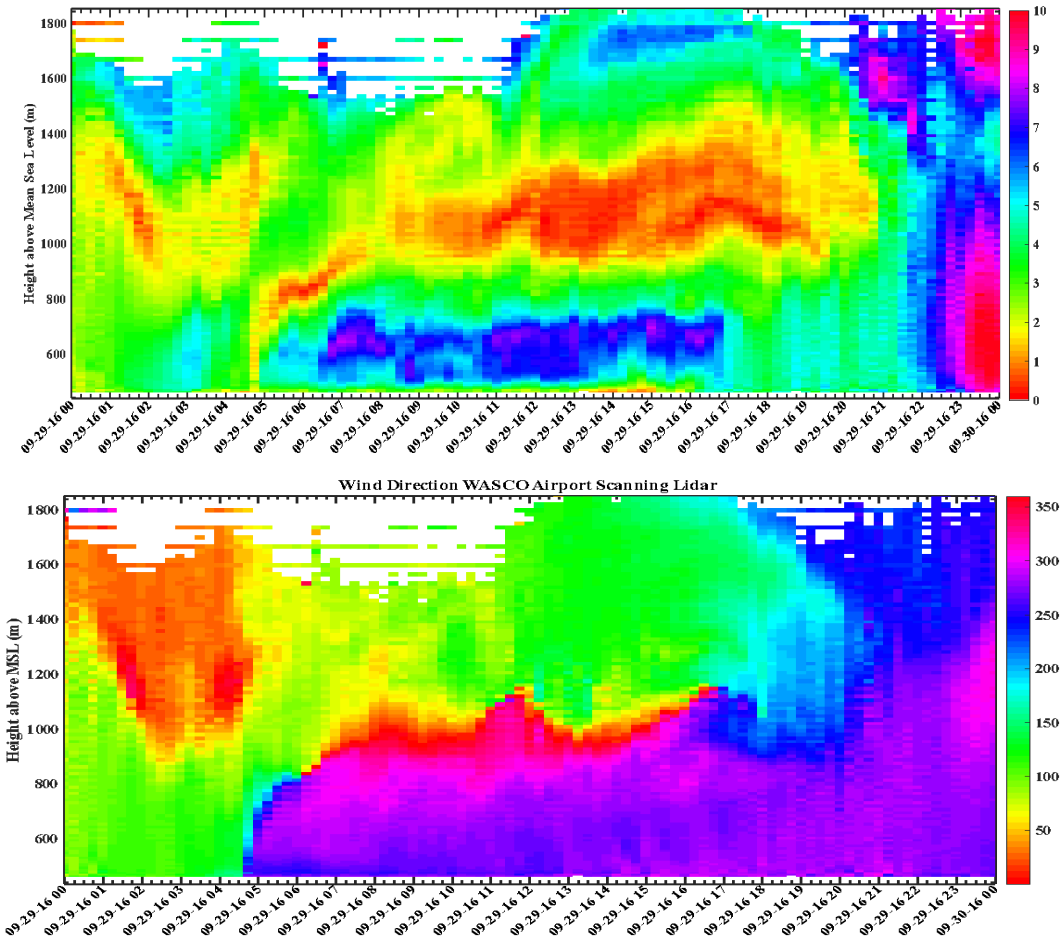

Figure 42. Horizontal wind speed in m/s (top) and wind direction (bottom) from scanning Doppler Lidar measurements at the WASCO site. The measurements are averaged over 15 minutes. Time in UTC.

iii. Case Study III — Dividing Streamlines on January 21st 2017

During stable wind conditions, the relationship between the Froude number Fr and the structure of flow encountering an obstacle is interesting. When Fr < 1, Sheppard (1956) argued that the flow would go around an obstacle. Hunt et al. (1997) and others extended Sheppard's work for prediction of stable flow structure in complex terrain. PS02 and PS06 tower measurements were used in the following analysis.

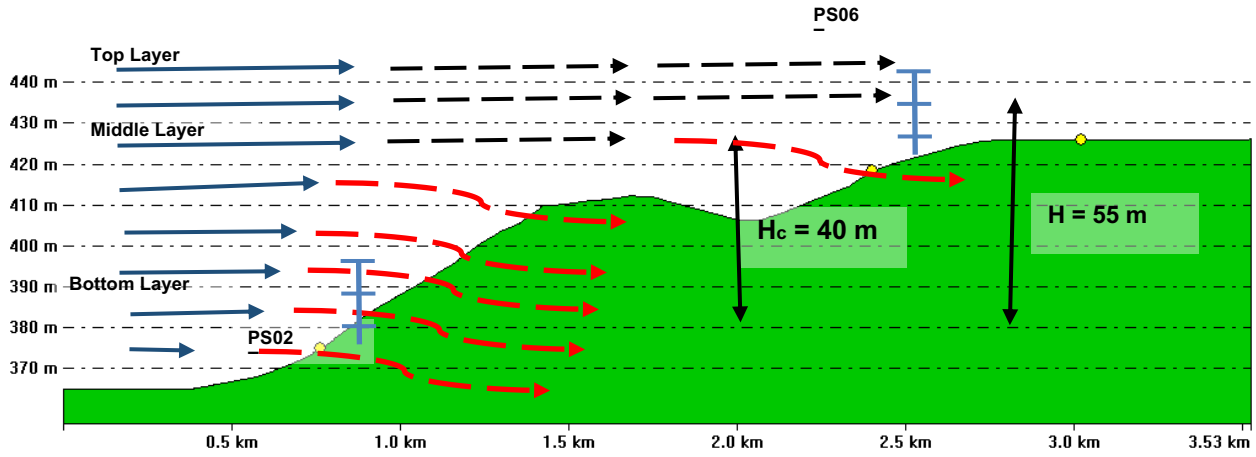

Figure 43. Rough illustration of dividing stream-lines observed at the Physics Site. The critical height (Hc) calculated based on Shepard et al., 1956 was approximately 40 m.

The Froude number was calculated from the two tower measurements according to

$$Fr = \left[\frac{U^2}{-\frac{g}{\rho}\frac{\partial \rho}{\partial z}(\Delta H)^2} \right]^{\frac{1}{2}}$$

where, ρ is the density of air, g the gravitational acceleration, $\frac{\partial \rho}{\partial z}$ is the appropriate density gradient, U is the free stream velocity at the top of the mountain and ΔH is the height to the top of the terrain. The classical critical dividing streamline height (H_c), at which height the potential energy is small compared to the kinetic energy of the flow, and whence the flow can go over the hill, was calculated using the Shepard et al., (1956) equation,

$$H_c = H\,(1 - Fr).$$

An illustration of the dividing streamline concept for the Physics Site is shown in Figure 43. The Froude number calculation was performed for several days, and an interesting case study is shown below.

On January 21, 2017, the winds were predominantly observed to be from the West. At 12:00 UTC [0400 Local time] a wind-direction shift was observed at 478 m above MSL at the PS06 tower location. Prior to the shift, the mean upwind wind direction was from the west ~ 270° . During the shift, the Froude number dropped below 1 (as seen in Figure 44 A). During low wind speeds, fluid parcels have less kinetic energy to go up the hill, and thus would flow around the hill (Hunt et al. 1997). The height from the flat terrain upwind of the hill to the hill top (H = 496 m) was 69 m, hence the critical height would be less than 69 m when Froude number is less than 1. Since there was no visual proof (such as smoke release) that the

flow went around hill, the wind direction at towers PS02 and PS06 were analyzed to investigate possible lateral flow distortion as the flow arrives at the hill. Figure 45 shows that similar wind directions prevail at the two upwind sensors on PS02, up to a height of 467 m above the MSL. When Fr dropped below 1, up to 50o shift in the wind direction was observed at PS06 (3m), which indicates possible wind deflection near the ground to flow around the hill. The wind direction at the top of the hill (i.e., level 21m sensor at PS06) approximately matched the upwind wind direction up to 1500 hrs. UTC (0700 hrs. LT), and thereafter the wind directions at the top of the hill deviated from the upwind wind direction as well. The critical height was calculated to be approximately 30 - 40 m above the upwind flat terrain height (Figure 44 C). Further investigation is needed to confirm above inferences.

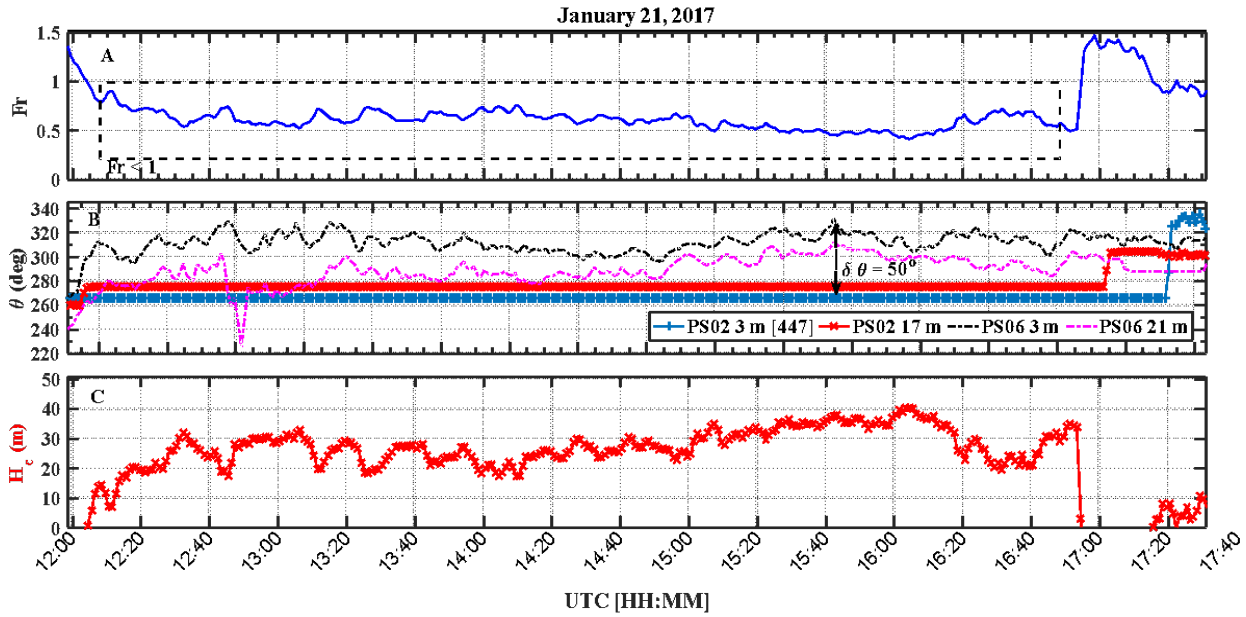

Figure 44. A) Froude number calculated from tower measurements on January 21, 2017 from PS02, B) time series of wind directions observed at various height above mean sea level and C) critical height (Hc) calculated as per Shepard et al., 1956, based on Froude number and height difference from the flat terrain upwind of the hill.

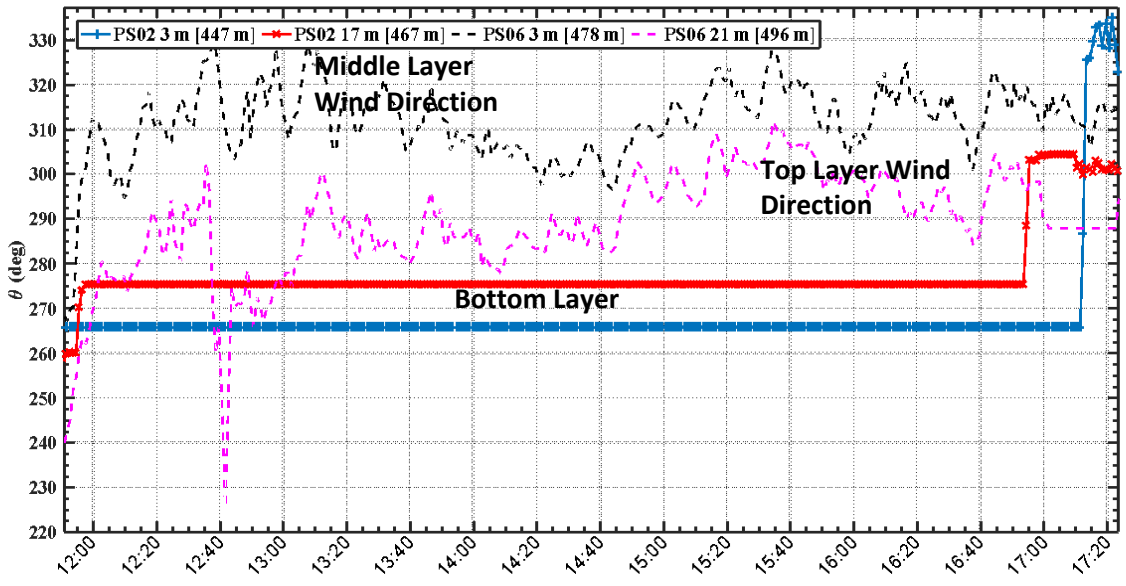

Figure 45. Wind Directions vs UTC time observed at various height above a terrain following surface from tower measurements. The legend also shows the height above mean-sea-level in squared brackets.

Further in-depth analysis of similar cases needs to be performed to study the effect of low Froude number effect on wind turbine loading, performance and forecasting.

3.3.3 Uncertainty Quantification

As part of the larger uncertainty quantification (UQ) effort, the Texas Tech team investigated the parametric sensitivity of wind speed forecasts in the context of an ensemble Kalman filter (EnKF) ensemble during high impact wind ramp events. This study built on the UQ work done by the team at Pacific Northwest National Laboratory (PNNL) described in Yang et al. (2017) and uses a similar UQ methodology. Based on the results of Yang et al. (2017), this study investigated a reduced set of 9 PBL and surface scheme parameters, shown in Table 15, during from two ramp events; a stable mix-out false alarm from January 18, 2017 and a marine push event from July 22, 2016. The primary goals of this work were to examine whether boundary layer scheme parametric sensitivity varies across different equally-likely ensemble members, different ramp-producing forecast situations, different horizontal grid spacings, and different forecast times.

Table 15. List of MYNN parameters to be modified in the parametric sensitivity study. The default values in the 2.5 level scheme in WRF 3.6 are listed as well as the range used in the sensitivity study.

Parameter	Symbol	Default Value	Range
Constant in TKE dissipation	B1	24	12-36
Prandtl number	Pr	0.34	1.5-4.5
Constant in LT calculation	$\alpha 1$	0.23	0.115-0.345
Constant in Ls calculation	$\alpha 4$	20	20-100
Constant in Ls calculation	$\alpha 5$	2.1	1.35-4.05
Exponent in Ls calculation	β	0.2	0.1-0.3
Surface roughness scaling factor	zf	1	1-2
Van Karman constant	k	0.4	0.35-0.40
Closure constant	$\gamma 1$		

Time-height plots showing SODAR, profiling radar and microwave radiometer observations for the January 18 case are shown in Figure 46. The mixing of warm, high momentum air towards the surface can be seen in the radar and radiometer measurements. Although the high momentum air did not mix down to the turbine rotor layer, the event log team identified this as an important case since over mixing by the model produced errors in forecasted wind speed.

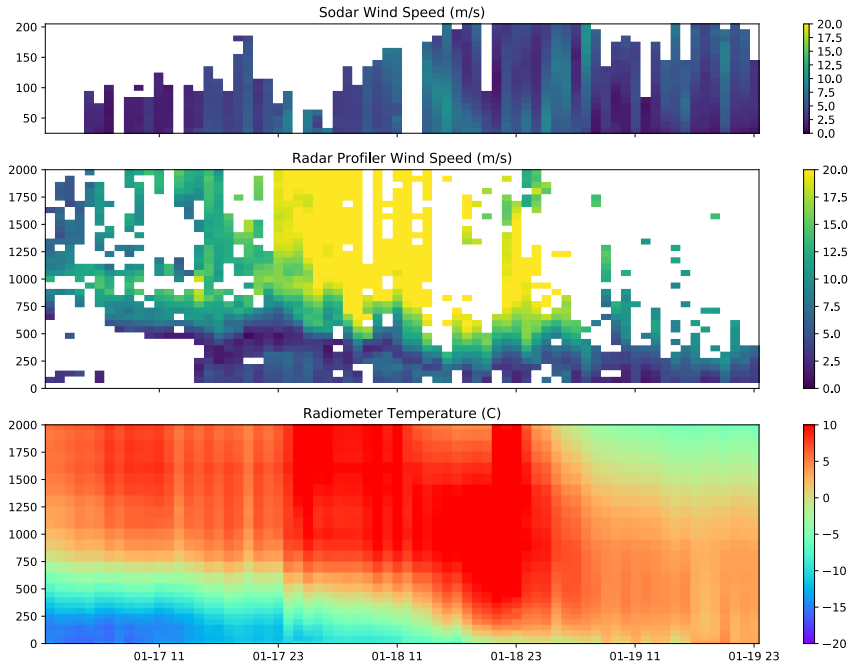

Figure 46. Time height plots of wind speed and temperature at the Wasco Airport for January 18, 2017 event. Vertical axis for all plots is in meters above ground level.

Figure 47 shows the average 80-meter above ground level (AGL) wind speed at Wasco for each member of the EnKF ensemble. The dashed lines show SODAR and radar observations from the Wasco airport valid at 80 and 81 meters AGL, respectively. Since each member of the EnKF ensemble is run with the standard 2.5-order MYNN scheme, all variation in this ensemble is a result of differences in the initial conditions (ICs) and lateral boundary conditions (LBCs). Three members of the EnKF ensemble were selected to provide ICs and LBCs for physics ensembles. The three members are shown in Figure 47 by the bolded green (high wind speed member), yellow (moderate wind speed member) and red (low wind speed member) lines. The members were chosen to represent different portions of the ensemble distribution.

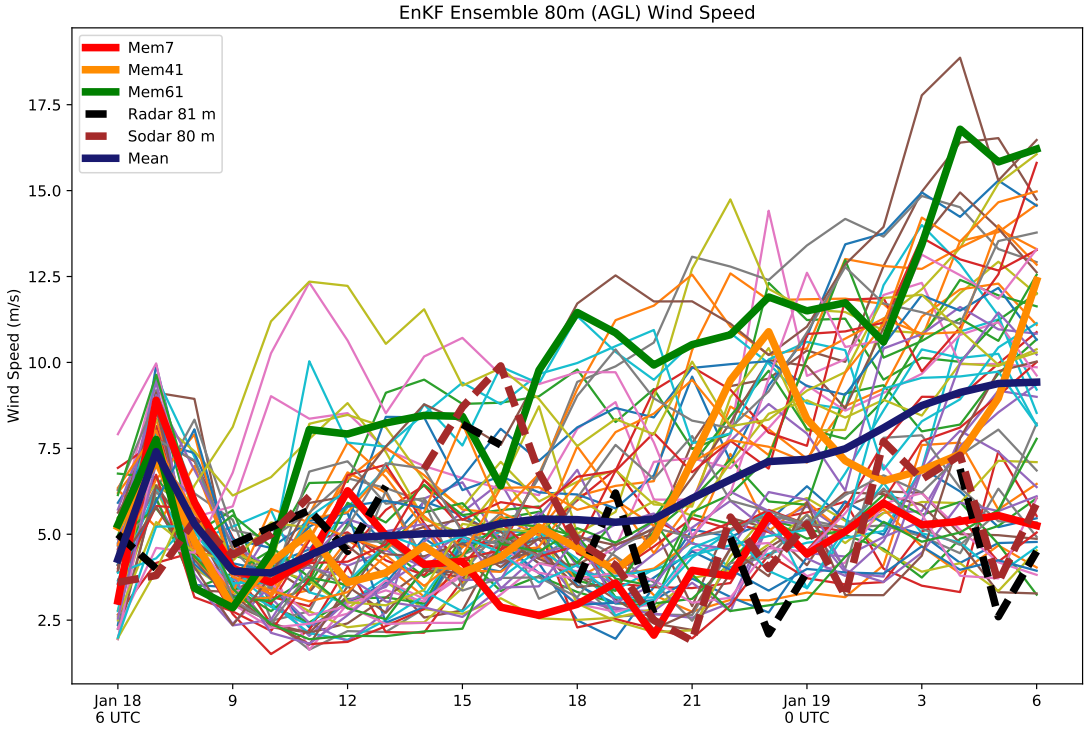

Figure 47. Time series of 80-meter wind speed at the Wasco airport for all members of the EnKF ensemble. The fast (green), medium (yellow) and slow (red) wind speed members that will be used for UQ experiments are bolded as is the ensemble mean (dark blue). The dashed lines show 80 and 81 meter AGL wind speed observations from a SODAR and profiling radar, respectively, deployed at the Wasco airport.

Each physics ensemble consists of 81 members, with all members using identical ICs and LBCs but a unique combination of MYNN parameters. Figure 48 shows the forecasted wind speed at Wasco for the three physics ensembles. Members of the physics ensemble forced by the low wind speed ICs, top panel, are quite similar to each other, suggesting the varied parameters have little influence on the modeled wind speed. Figure 49 shows the variance in wind speed for all three physics ensembles as well as the wind speed variance from the EnKF ensemble. The larger variance in the EnKF ensemble, when compared to the physics ensembles, suggests that forecast uncertainty for this case is dominated by IC error rather than errors in the MYNN scheme

Following the method used in Yang et al. (2017), the wind speed variance for each physics ensemble was decomposed using a generalized linear model (GLM) into the portions contributed by each parameter. Figure 50 shows the percentage of total variance than can be attributed to each parameter every hour. For all three ensembles, B1, $\gamma 1$ and the Prandtl number are the dominant sources

of forecast uncertainty, however, there are key differences between the ensembles. In the low wind speed ensemble B1 is responsible for the majority of the forecast variance during two periods between 13-14 UTC and 17-19 UTC on January 18. Between these periods, $\gamma 1$ accounts for more variance than any other parameter, while Pr accounts for the most variance early in the forecast and between 22-23 UTC on January 18. During the final 6 hours of the forecast, no parameter is responsible for more than 20% of the total variance. The medium wind speed ensemble sees B1, $\gamma 1$ and Pr alternate as the most influential parameter for the first 18 hours of the forecast though no parameter accounts for the majority. During the final six hours of the forecast B1 is responsible for the majority of the variance in forecast wind speed. In the high wind speed ensemble, as in the medium wind speed ensemble, B1 is the single most important parameter during the final six hours of the forecast. However, in the high wind speed ensemble B1 is also the most important parameter between 11 and 18 UTC on January 18.

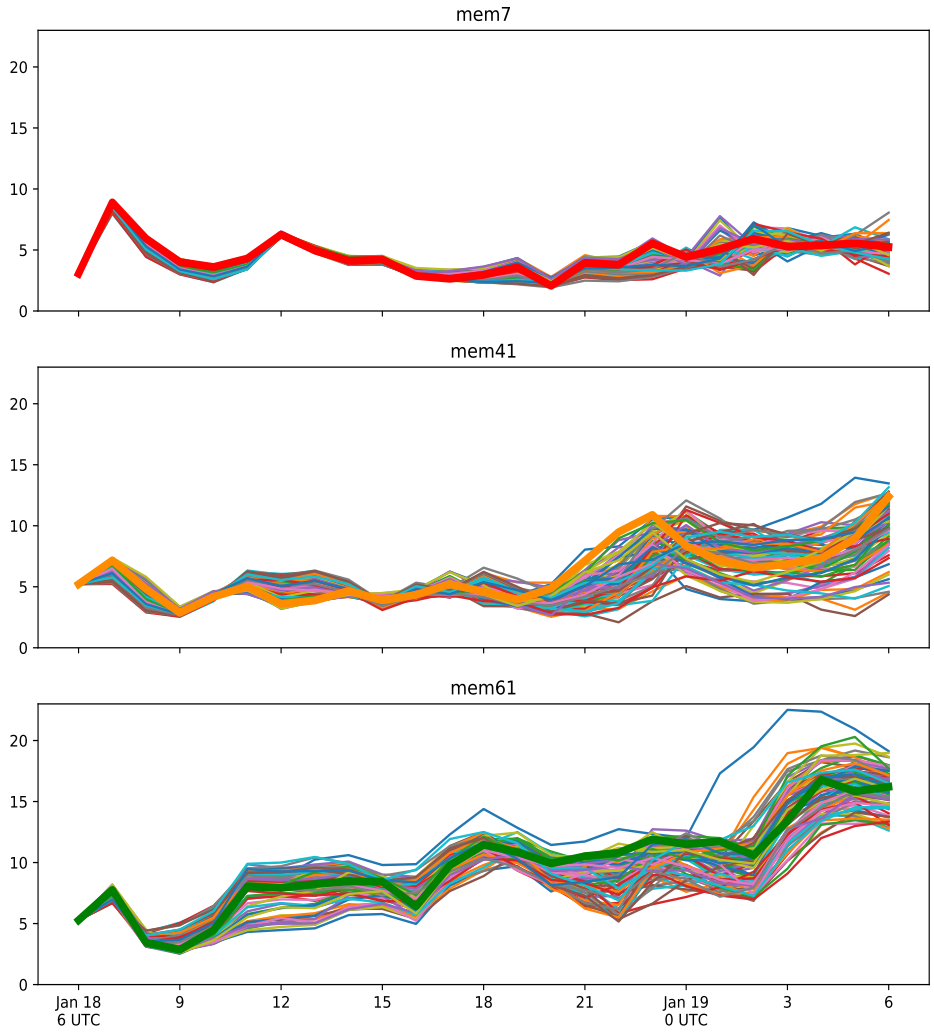

Figure 48. Time series of 80-meter wind speed at the Wasco airport for the three physics ensembles for the January 18 case. The bolded lines are the EnKF members from Figure 2, which provide the ICs and LBCs for their respective ensembles and are included here for reference.

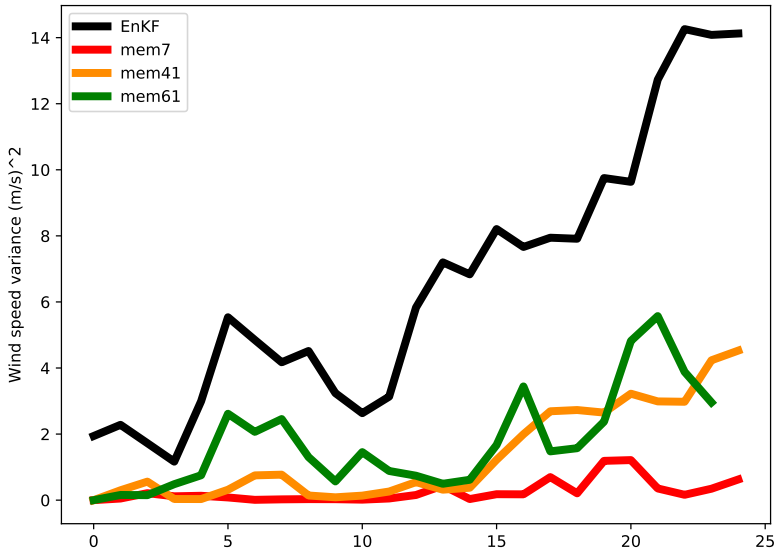

Figure 49. Variance in wind speed for the EnKF (black line) and physics ensembles (colored lines) for the January 18 case.

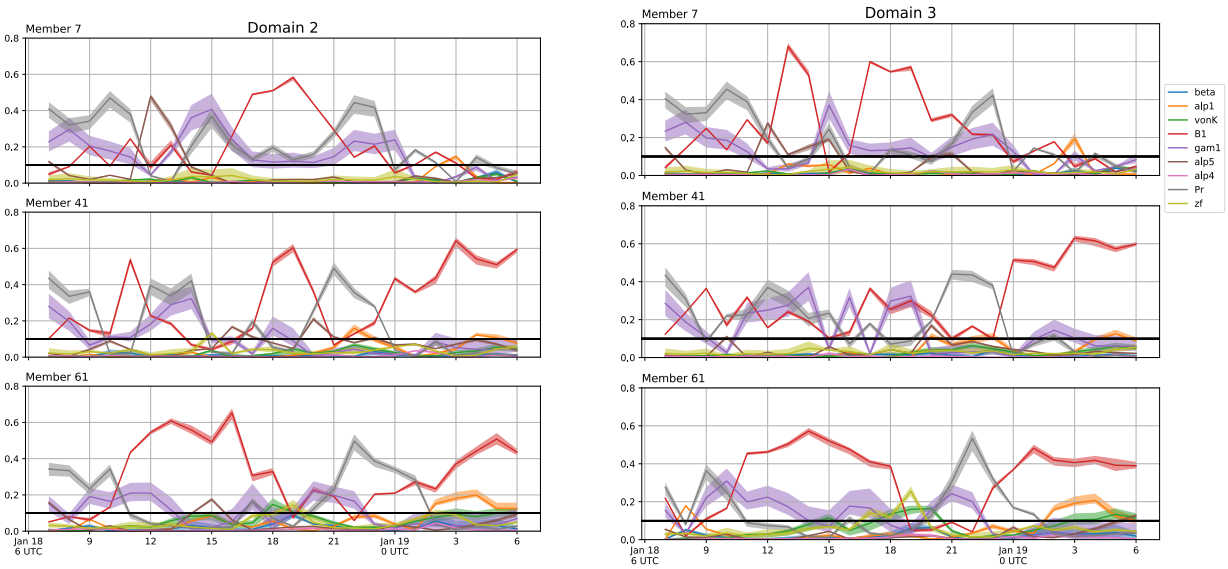

Figure 50. UQ results the January 18 case on Domain 2 (left) and Domain 3 (right) for the slow wind speed (top row), middle wind speed (middle row) and high wind speed (bottom row) runs. The lines represent the percentage of wind speed variance that can be attributed to each parameter. The vertical axis units are percentage of variance. The horizontal black line indicates 10% of the total variance.

The difference in UQ results between these ensembles highlights the importance of addressing IC uncertainty when conducting sensitivity experiments and especially when tuning model parameters. Each of these members is an equally likely draw from a distribution of possible states. Studies done using a deterministic forecast are essentially selecting one of these possibilities and are unable to assess the other, equally likely, initial states. Tuning the model parameters based on a single set of ICs ignores a significant source of model error and can potentially introduce more error by modifying model physics that may not be at fault.

The left column of Figure 50 shows UQ results from Domain 2. The UQ results between Domain 2 and Domain 3 are quite similar, which suggests that the 4 km grid spacing is sufficient for future UQ work.

Figure 51 shows the average 80-meter wind speeds at Wasco for the EnKF ensemble for the July 21, 2016 marine push event. For this case, 5 members were selected to drive physics ensembles. Three members that captured the event and produced large (green), medium (yellow) and small (red) wind ramps were selected along with a member that failed to capture the event (brown). A final member (blue) was selected as it has an odd surface pressure signature (not shown). The wind speed time series for each of the physics members is shown in Figure 52. The varied physics parameters have very little impact on the low wind ramp and bust ensembles and produce moderate variation in the large ramp case. In contrast, the PBL scheme is responsible for large variations in wind speed in the moderate ramp ensemble and for variations in the timing of the ramp event in the odd pressure ensemble. The variance for each ensemble is shown in Figure 53. Unlike the January 18 case, two of the physics ensembles produce comparable forecast variance to the EnKF ensemble. However three of the physics ensembles produce very little variance.

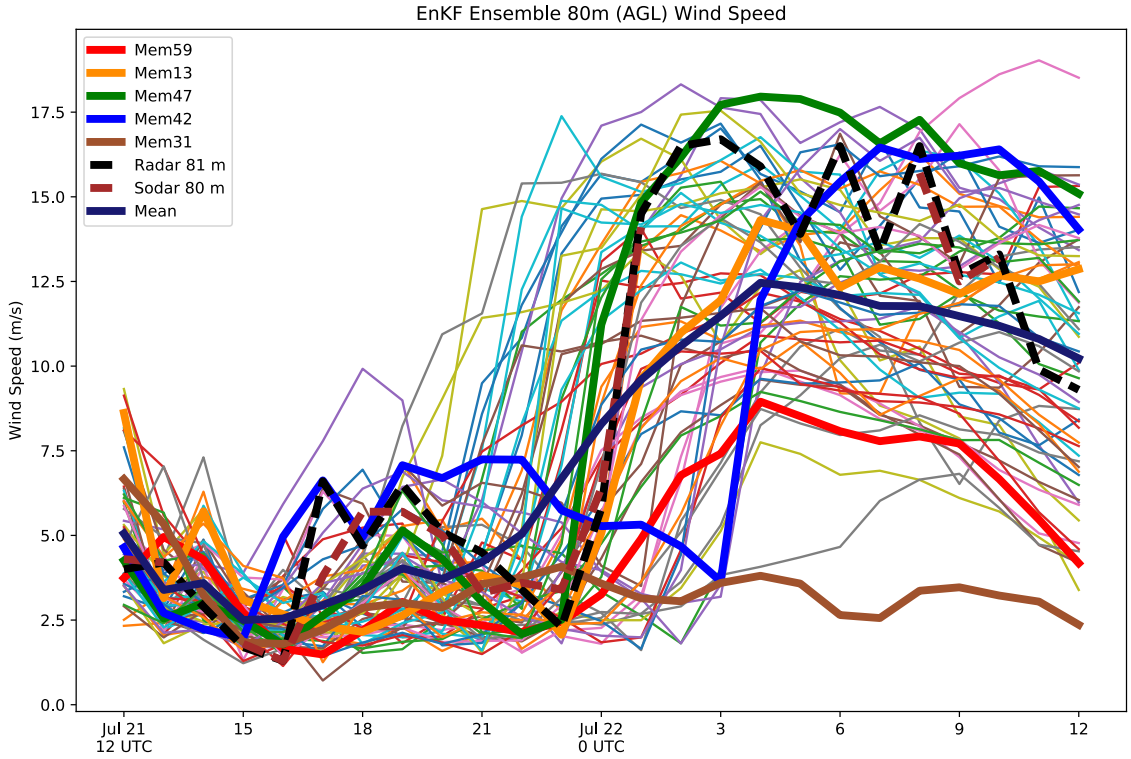

Figure 51. As in Figure 47 but for the July 22, 2016 case.

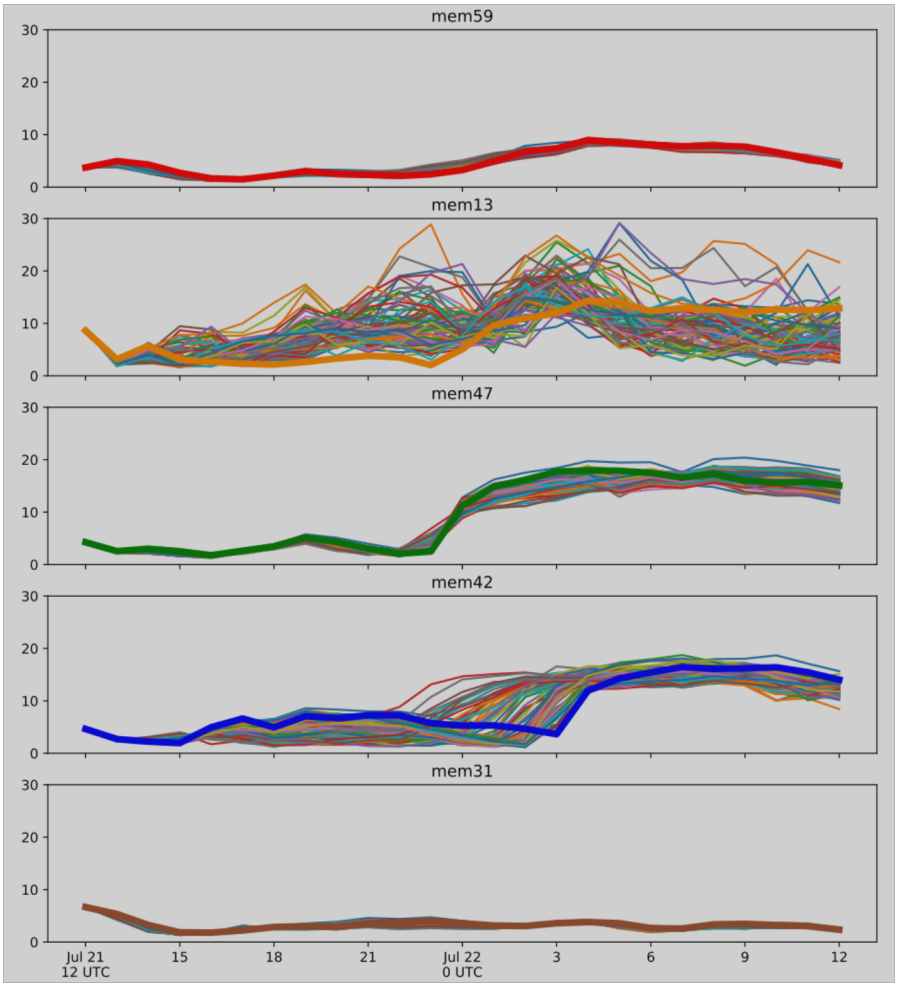

Figure 52. As in Figure 48 but for the July 22, 2016 case.

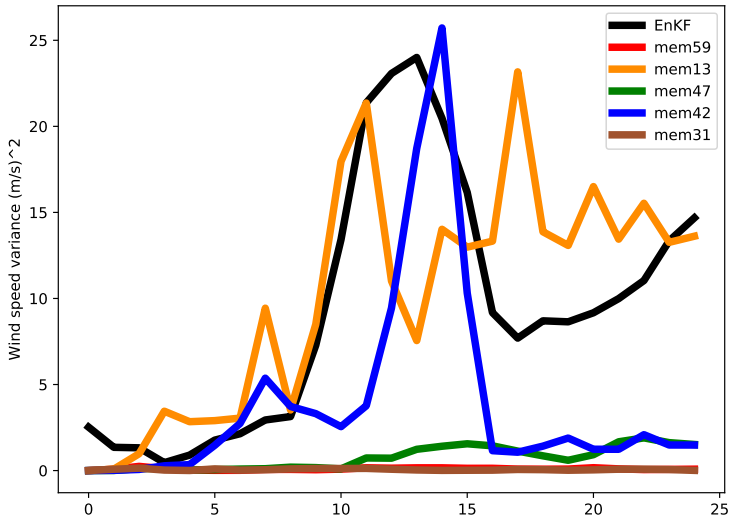

Figure 53. As in Figure 49 but for the July 22, 2016 case.

Figure 54 shows the UQ results for the July 22 case. The sensitivity to individual parameters varies widely between the different ensembles, but the results from domain 2 and domain 3 are fairly similar. As with the January 18 case, the difference in parametric sensitivity between the ensembles highlights the fact that variations in ICs can influence the physics parameterizations as well as the difficulty in isolating physics error when dealing with real world forecasts. Furthermore, the fact that the variance from the EnKF ensemble exceeds the variance of all the physics ensembles for the January case and most of the physics ensembles from the July case suggests that forecast uncertainty in these events is largely due to IC error and not errors in the PBL parameterization.

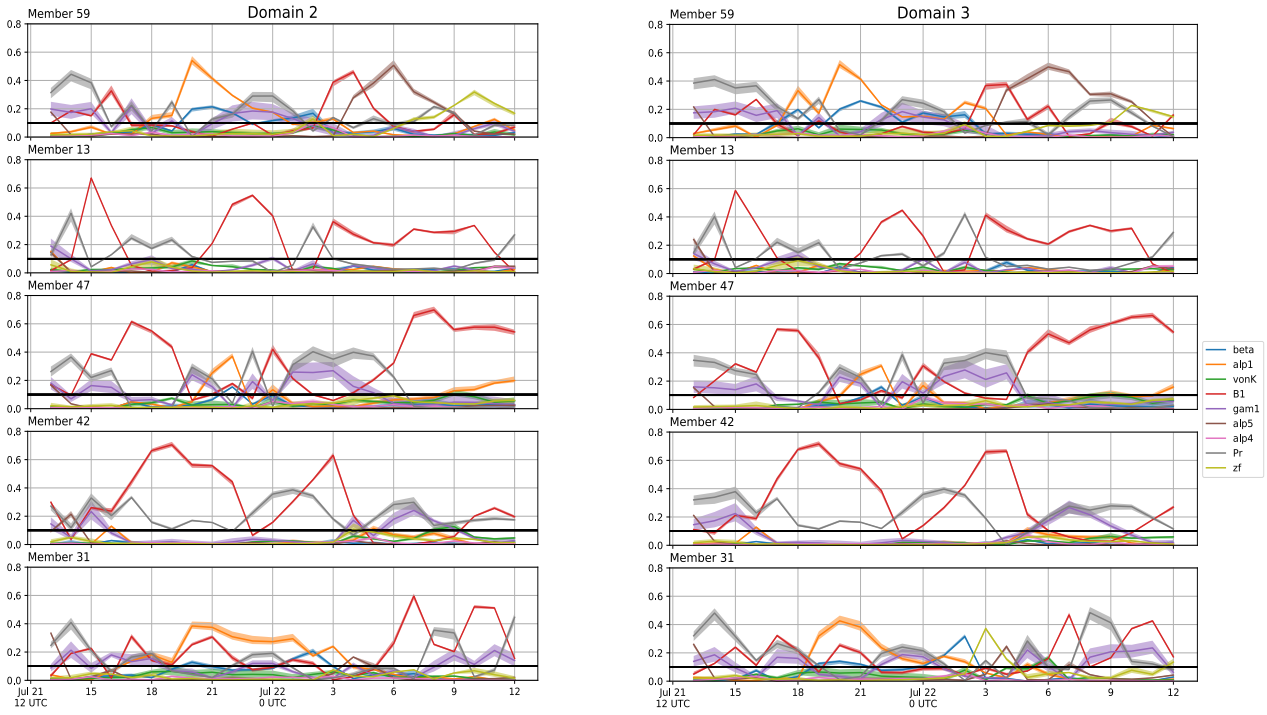

Figure 54. As in Figure 50 but for the July 22, 2016 case.

3.3.4 Wind Power Forecasts

A part of the Vaisala team project objectives within WFIP2 was to generate and validate wind power forecasts based on the control (baseline) and experimental (enhanced) HRRR reforecast simulations produced by NOAA. The original project plan aimed to produce retrospective forecasts for every major wind project connected to the BPA transmission system for a minimum of one calendar year period, spanning all four seasons during the WFIP2 field campaign. Due to data restrictions, not all of the facilities could be individually modeled. The Vaisala team, with the cooperation of its data partners, contributed data for 20 individual projects out of the total of 31 wind facilities within the BPA control area. As a compromise and as a way to measure performance for the entire installed wind capacity of the area, the total regional power generation was utilized, since it is freely available and published in near real-time on the BPA website[1]. Although the WFIP2 reforecast simulations were eventually created for one full year of the WFIP2 study period, at the time of this analysis, only four 1-month periods were available for download. These months were April, July, and October of 2016 and January of 2017,

[1] https://transmission.bpa.gov/Business/Operations/Wind/twndbspt.aspx

calendar months which are centered during each of the four seasons. This was considered to be the minimum necessary to proceed with a seasonal analysis of the potential benefits of the WFIP2 modeling improvements to wind power forecasting, but the study is still considered to be preliminary because of the limited sample size. Regardless, we anticipated seeing more subtle improvements to the wind power forecasts than to the atmospheric model outputs because of the statistical post-processing involved. As illustrated in Figure 55, statistical post-processing acts to reduce the systematic error of raw model output and is typically more effective at larger starting error levels.

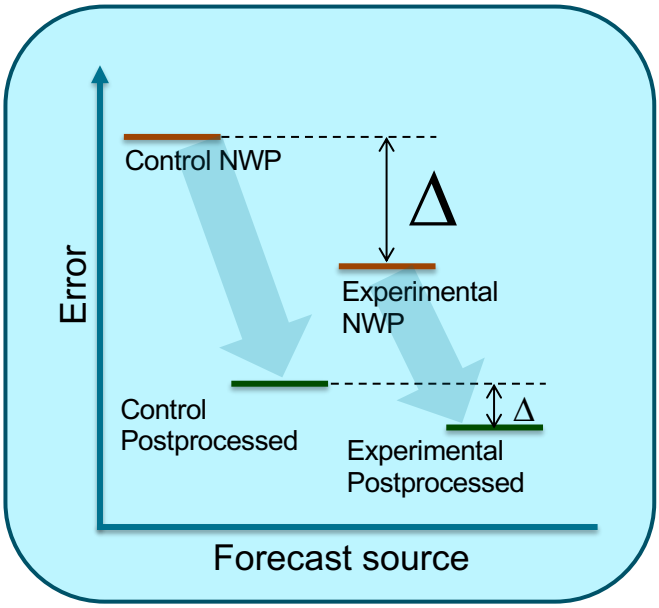

Figure 55. A general schematic of the potential reduction of average forecast error by a statistical post-processing method starting from a higher-error NWP model versus an improved experimental NWP model. There is less inherent forecast error in the experimental NWP model and therefore the statistical correction removes less overall error than it does from the control model. The relative delta in forecast error between the control and experimental NWP model output is larger than the difference between the post-processed outputs of each.

3.3.4.1 Methodology

Vaisala used the provisional WFIP2 HRRR reforecasts produced by NOAA as foundational weather model inputs to drive statistically corrected forecasts of wind power at 1-hourly intervals out to 24-hour horizons. Separate wind power forecasts were generated for the control and experimental model configurations so that we could inter-compare their performance overall and during specific cases of interest. Actual power generation data served as the target variable for the forecasts and as the source of validation data. Performance metrics were kept to a very basic set within this evaluation and we leave it to future studies to assess the potential economic impacts of the improved HRRR model from WFIP2.

i. Data

Substation level generation and project available capacity data was provided by Vaisala team data partners for 20 participating individual wind projects. Although some contributed data at the wind turbine unit level, this more detailed information was not directly used in the forecast generation process. Turbine level data was used in some cases to determine the project-wide available power capacity. The actual power divided by the available power capacity, or relative power, served as the primary predictand for our wind power forecast modeling work. In the case of the regional forecast for the BPA fleet aggregate, we used the actual wind power reported by BPA and normalized by the installed wind

capacity of the region (4783 MW) to obtain a pseudo relative power. All relative power data was converted to 1-hour period-ending averages to correspond with hourly wind power forecast intervals as set out in the aim of the study. We did not attempt to analyze sub-hourly forecast accuracy or variance.

As described above, for NWP data we used the provisional WFIP2 HRRR reforecasts available from the Data Archive and Portal (https://a2e.energy.gove/about/dap). These simulations are "cold-start" configurations with initial conditions taken from the RAP and run twice per day (every 12 hours) through forecast hour 24. Otherwise, the WFIP2 HRRR (Δx = 3 km) utilizes the same domain setup and physics suite as its operational counterpart. A very high-resolution nested domain, nicknamed HRRRNEST, is run at Δx = 750 m inside the WFIP2 HRRR with a delayed start of 3 hours. We examined only the HRRRNEST runs in this study due in part to initial delays in accessing this large volume of archives from the WFIP2 Data Archive and Portal. The HRRRNEST runs target the terra incognita scales (200 m < Δx < 1000 m), where PBL parameterizations are known to struggle, but better resolve important terrain features of the region including the Columbia River Gorge. At the time of analysis, the HRRRNEST reforecast data for April, July, and October of 2016 and January of 2017 was available to us. A more complete analysis of both WFIP2 HRRR and HRRRNEST simulations for the full year of available reforecast data is now possible.

To obtain NWP-based predictors at each location, the HRRRNEST model output was bi-linearly interpolated to each wind project centroid from the corresponding four surrounding grid points. No time interpolation was necessary as we used only hourly model output to match up with our hourly resolution power data. We sliced out forecast hours 4 through 15 from each model run cycle, each being 12 hours apart, in order to stitch together a single complete training data set for all available valid times within the four one-month reforecast periods. A large variety of standard model output variables from the HRRRNEST runs are available in the WFIP2 Data Archive and Portal. We utilized all of the 2-D fields as well as the wind 3-D fields on pressure surfaces. Additionally, we derived the wind speed from its vector components, but included all three wind related variables in the combined training data set to have both magnitude and zonal/meridional components available as candidate predictors.

ii. Modeling Approach

Statistical wind power models were developed on the BPA fleet aggregate and on the individual site level. Statistical models were also created for each hourly forecast horizon independently. No attempt was made to forecast wind speed (either hub-height or rotor-equivalent) and then translate to power output with a power conversion model. Our approach is to forecast the relative power directly. Vaisala used its in-house wind power forecast engine to train the statistical models, which is part of its standard operational software.

The Vaisala power forecast engine is built around the foundation of open source machine learning software, so that any widely used statistical model type can be deployed for a particular target variable and forecast time horizon. The system automatically selects the most appropriate components from the full collection of input data, while controlling for redundant information from similar variables using a sophisticated feature ranking and selection technique. An efficient grid search is used to tune the statistical model hyper-parameters and a k-fold cross-validation strategy is employed to avoid over-fitting to the training data. Commonly chosen algorithms include support vector regression (SVR), gradient boosted regression (GBR) trees, and ridge regression (RR), though a wealth of other statistical models are available should they prove to generalize better for a particular location or time period. For 0-6 hour forecast horizons, recent observations from the wind facility have great importance and are weighted heavily by the chosen predictor variables. Nearby weather observations, taken from off-site locations, can lead to further improvements. Even though a rich network of off-site observations were available during WFIP2, we did not include them as candidate predictors in this study choosing instead to focus on the NWP model improvements. Beyond about the 6-hour forecast horizon, recent observations tend to provide little or no tangible benefit to power forecast accuracy, so at these intervals, the predictors coming from the NWP models dominate. Because of the similarity in relative importance of model predictors beyond forecast hour 6, we chose to pool forecast hours 7-15 and 16-24 into groups and

thereby lessen the total number of models we had to build. This reduced the number of horizon-specific models from 24 down to 8 for each location and model type.

Due to limited sample size of the data period in the reforecast study, with only a one-month sample from each season available, we could not afford to split the available periods into fixed training and testing sets. Instead we employed a k-fold cross-validation approach to assess how the results of our models would generalize to independent data. We divided the data set into eight folds, each one of approximately a half-month in length, leaving out one half-month partition as testing data and using the remaining 3.5 months as training data. We rotated through the folds, using each half-month period as a validation set, as pictured in Figure 56. Days 1-15 were separated from days 16-end of the month. This had the benefit of including half a month of in-season data within the training data set when predicting the other half-month from the same season. With calendar month splits only using 4 folds, it would not have been possible to include in-season training data for the target periods with the available data set.

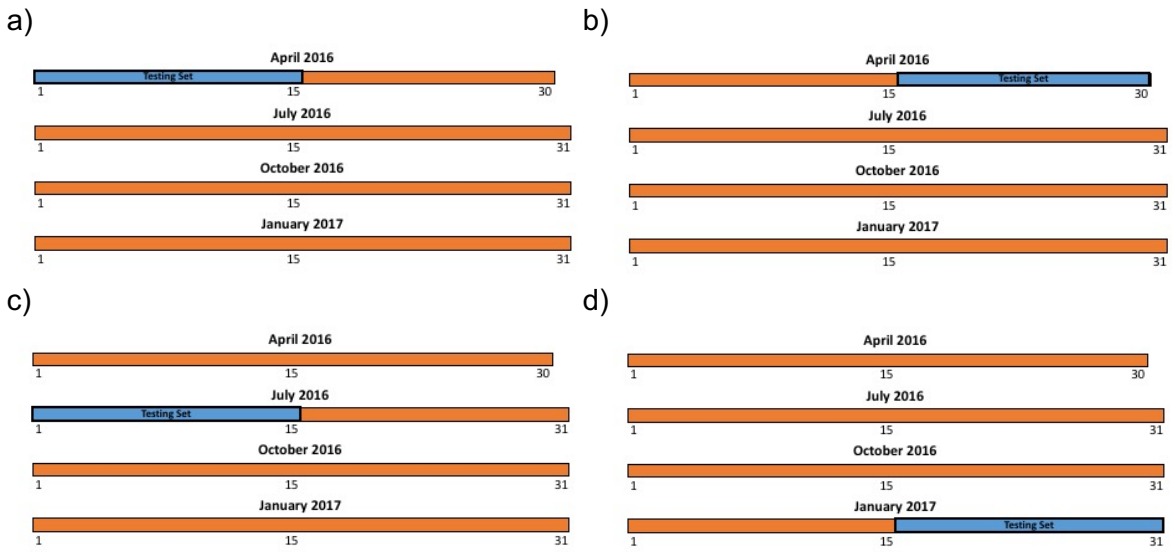

Figure 56. Illustration of the k-fold cross-validation wind power forecast training strategy with successive half-month periods withheld as test data (blue), leaving the remaining dates available as training data (orange). The panels show a) the first fold using.

The total number of statistical forecast models built for each site was 128, taking 8 x 8 = 64 models each for the control and experimental HRRRNEST runs across each horizon and training fold. For 20 individual wind projects and the single BPA aggregate, that meant we trained a grand total of 21 x 128 = 2,688 statistical models. Each model was saved and the wind power forecast output from each half-month test period was organized. A script was written to merge the output from each training fold into a single time series file, organized by forecast horizon. All wind power reforecast files were delivered to the WFIP2 Data Archive and Portal.

3.3.4.2 Forecasting Validation Results

The normalized MAE for HRRRNEST-based wind power forecasts on the BPA fleet aggregate are shown in Figure 57, organized by forecast horizon. Forecast hours 1-6 are scored individually, while forecast hours 7-15 and 16-24 are scored as pooled groups. Each month is scored separately, so that we can compare the performance by season, starting from winter (January 2017) in the upper left, and rotating through spring (April 2016), summer (July 2016), and finally fall (October 2016) in the lower right panel. Experimental model results are denoted in blue, while control is represented as red. Overall, the forecast accuracy difference is small between the control and experimental models, except in the winter

month. Normalized MAE ranges from about 3-5% at 1-hour forecast horizon up to 6-12% for the 18-24 hour forecast period. In Figure 58, we report the normalized MAE for BPA only on days within the WFIP2 Common Case Study Set. This represents about one third of the days in the overall reforecast period. The horizon-averaged MAE skill score, which represents the relative improvement by the experimental reforecasts over the control reforecasts at a summary level, are presented in Table 16 for both the Common Case Study Set days and the overall reforecast periods.

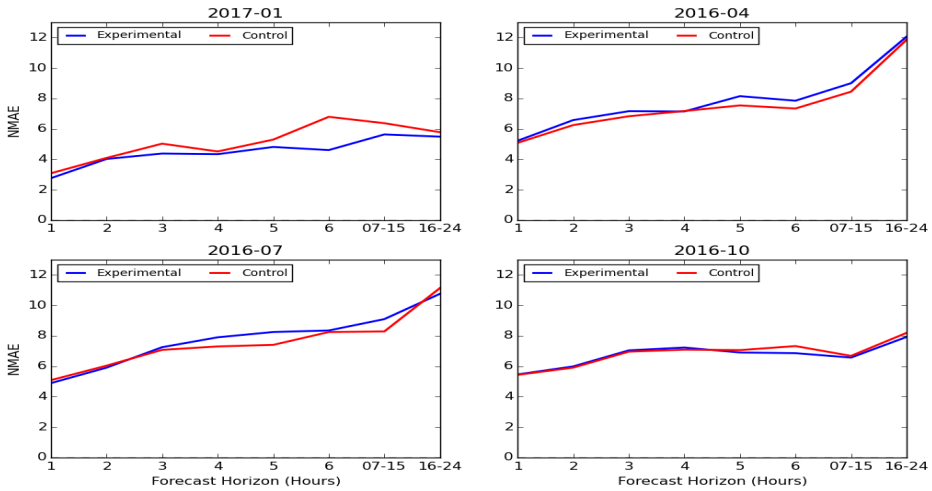

Figure 57. Normalized MAE for HRRRNEST-based wind power forecasts on the BPA fleet aggregate organized by forecast horizon and reforecast month. Control models are the red curves and experimental models are blue.

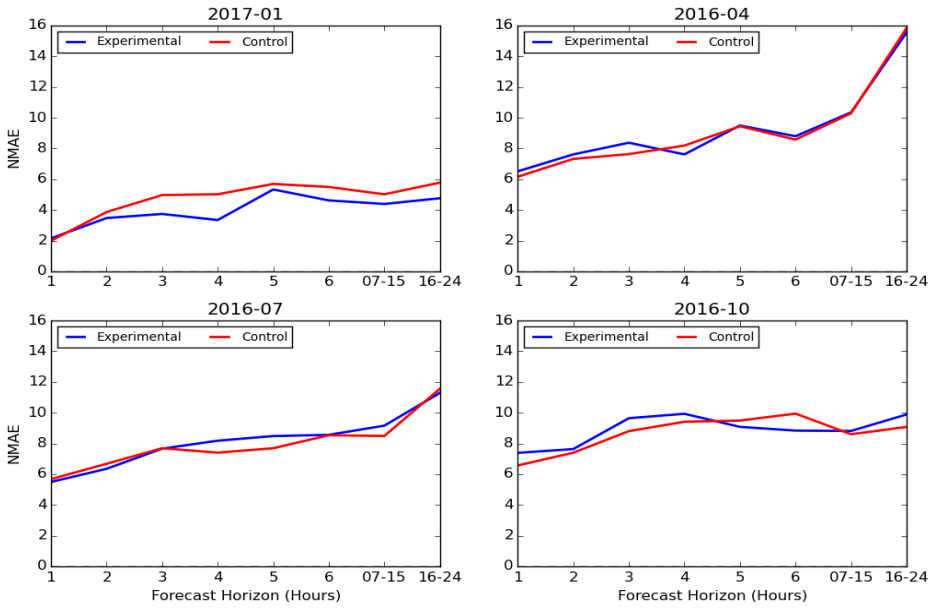

Figure 58. As in Figure 57, except for just days within the Common Case Study Set.

Table 16. Horizon-averaged MAE skill score for the BPA fleet aggregate, showing the improvement of the experimental HRRRNEST-based wind power forecasts relative to the control for the entire reforecast period and the Common Case Study Set.

Season	All Reforecast Cases	Common Case Study Set
Winter (2017-01)	11%	14%
Spring (2016-04)	-4%	-2%
Summer (2016-07)	-3%	-2%
Fall (2016-10)	1%	-3%

In terms of normalized bias, Figure 59 and Figure 60 show the results for BPA with the same organization by forecast horizon and season. We observe negative forecast biases in winter and spring, with an indication of small bias reductions by the experimental runs. For the Common Case Study Set days during the winter month of January 2017, the sign of the bias actually switches to positive. The experimental runs reduce this positive bias toward zero for 6 of the 8 forecast horizons. The overall negative biases are larger in the spring month and even reach below the -5% level during April 2016 days within the Common Case Study Set. The springtime reduction in negative bias by the experimental forecast averages 0.2% for all days and 0.6% for Common Case Study Set days. In summer, the forecast biases are positive and are not reduced by the experimental model, although they appear to be smaller on average for the Common Case Study Set days. The fall biases are negative, but relatively small, with absolute values below 2% overall. However, October 2016 days within the Common Case Study Set exhibit larger negative biases and no apparent improvements from the experimental model.

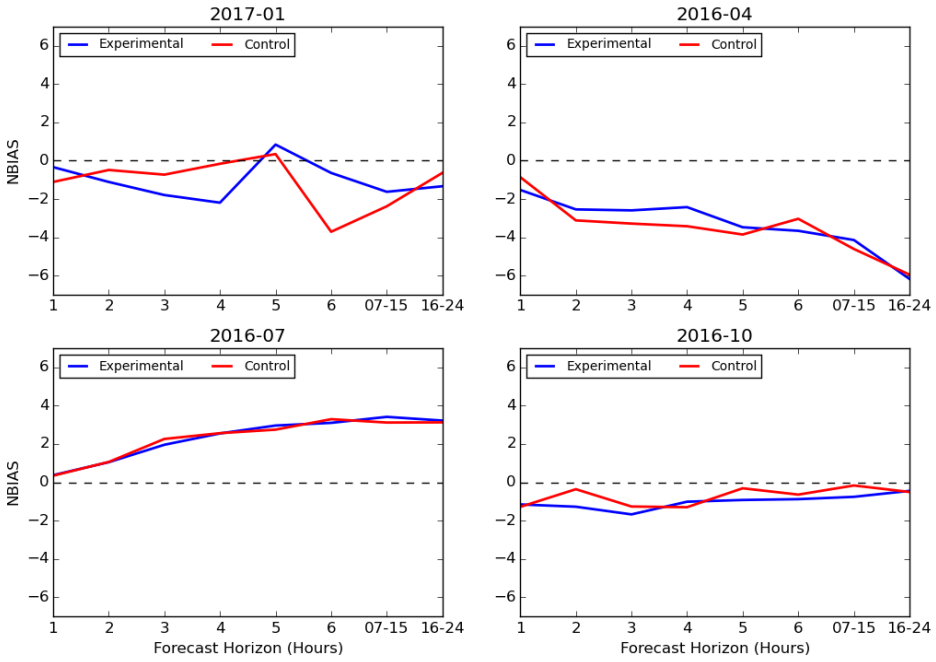

Figure 59. Normalized mean bias error for HRRRNEST-based wind power forecasts on the BPA fleet aggregate organized by forecast horizon and reforecast month. Control models are the red curves and experimental models are blue.

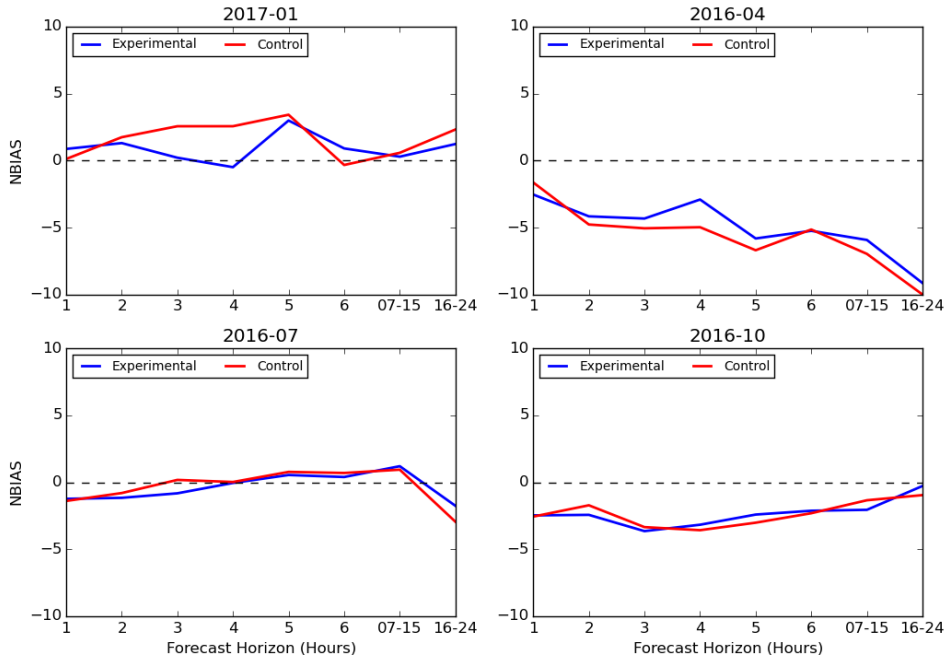

Figure 60. As in Figure 59, except for just days within the Common Case Study Set.

The same statistics are measured for wind power forecasts at the individual site level. In Figure 61 we show results for an indicative project located in the western part of the Columbia River Basin study area near the river gorge. Normalized MAE ranges from about 6-10% at 1-hour forecast horizon up to 12-16% for the 18-24 hour forecast period. We show the same performance measures in Figure 62 for a representative wind project in the Arlington, Oregon area where the forecast errors are generally lower and have greater seasonal variability. A summary of overall seasonal performance is included in Table 17 given by horizon-averaged MAE skill score and segregated by site.

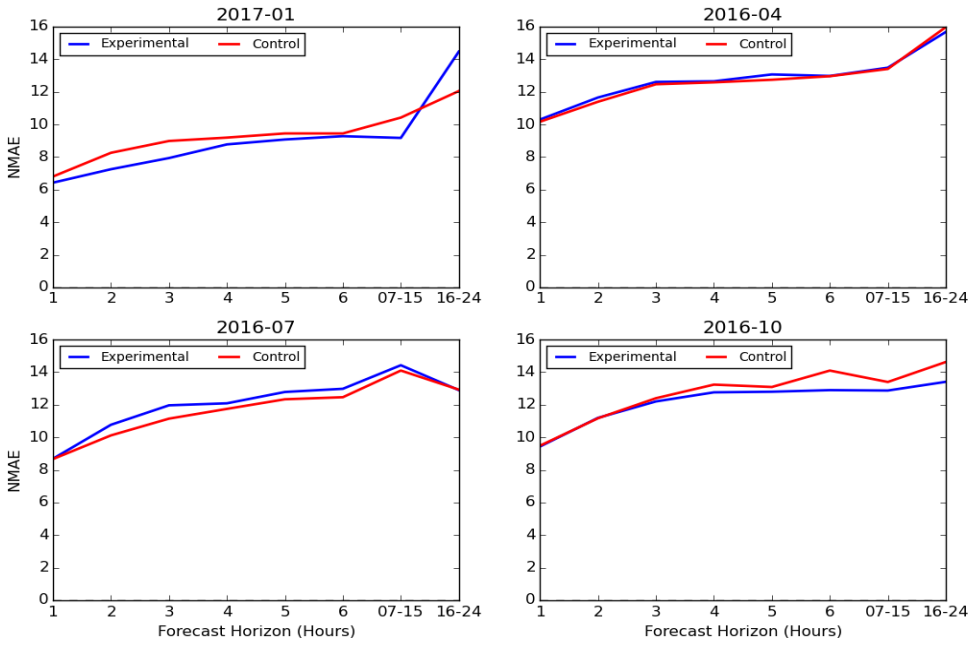

Figure 61. Normalized MAE for forecasts at White Creek Wind in Klickitat County, WA.

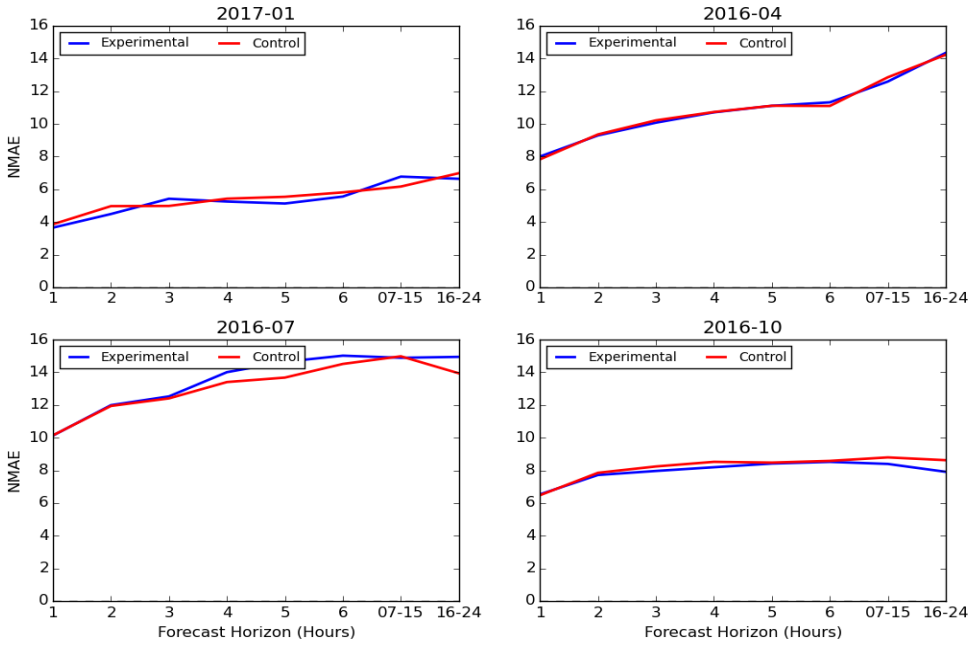

Figure 62. As in Figure 61, except for forecasts at North Hurlburt, near Arlington, OR.

Table 17. Normalized MAE skill score organized by individual wind project and by season.

Wind Project Name	Winter (January 2017)	Spring (April 2016)	Summer (July 2016)	Fall (October 2016)
Combine Hills II	-7.0	3.2	-3.4	4.4
Harvest Wind	3.0	-5.0	-0.6	2.9
Big Horn	-1.4	0.0	-1.0	-2.3
Big Horn II	-1.9	1.4	-1.6	0.1
Hay Canyon	-1.0	0.4	-3.9	-1.9
Juniper Canyon	-19.8	N/A	-1.6	-1.0
Klondike	3.2	2.8	-2.0	-1.1
Klondike II	-1.7	-0.5	-2.4	-0.5
Klondike III	4.0	1.5	-2.0	-1.7
Klondike IIIA	-2.0	3.0	-8.2	-0.3
Pebble Springs	-7.5	-3.1	-2.6	2.4
Star Point	-1.7	1.2	-1.3	0.8
Biglow I	N/A	7.7	-2.7	6.6
Biglow II	N/A	1.4	-1.2	4.4
Biglow III	N/A	3.4	1.6	3.9
Tucannon River	N/A	2.5	-0.1	-0.7
Horseshoe Bend	-1.5	-1.8	-4.0	1.5
North Hurlburt	2.0	-0.1	-2.9	2.8
South Hurlburt	3.2	-2.3	-4.0	0.0

3.3.4.3 Forecasting Validation Summary

Improvements in wind power forecasting provided by the experimental HRRRNEST model are mostly negligible in all months except winter. The overall improvement in normalized MAE for the BPA regional forecast in wintertime averages out to 11% for all horizons. For the subset of cases during January 2017 deemed particularly interesting to the project team and of potential importance for the wind industry, we achieve a slightly larger average improvement of 14%. However, all other seasons show no major benefits. Although a strong negative forecast bias seems to be reduced a little during the springtime, it does not translate to a noticeable reduction in the normalized MAE. On average, the results in the other three seasons even suggest some possible degradation. At individual wind sites, the wintertime forecast improvements are generally smaller than at the aggregated regional level, ranging from 0-5% in terms of normalized MAE. This is a typical result since the aggregated regional forecasts have a greater dependency on the meteorological patterns represented in a mesoscale model, while some of errors associated with microscale variations are averaged out in the regional forecasts.

In comparison to the raw model 80-m wind speed validation results presented by Olson et al. (2019), there are many similarities. For the HRRRNEST simulations over all four reforecast periods, the experimental runs reduced wind speed forecast errors by an average of about 2-7% over the 19 sodar locations distributed throughout the Columbia River Basin during WFIP2. The spring and summer periods showed the least impact. Whereas, improvements were the most robust in fall and winter, when the implemented physics changes that mostly benefited stable PBL regimes were more influential. While we did not realize any benefits during the fall period, the wintertime improvements approach the 15-20% level seen for the rotor-layer wind speeds. Our results for wind power are a little more pessimistic because they measure the improvements after statistical post-processing is applied, which acts to

reduce the differential between the experimental and control runs. Another factor contributing to our lower estimates of forecast improvement is the smaller sample size (four versus six week reforecast periods) and the need for longer training data periods overall to create more skillful statistical model corrections. Finally, the realized improvements should be larger at the operational HRRR resolution of 3-km, because some of the biggest physics improvements in WFIP2 came from parameterization schemes that were active only at those scales and not appropriate for the sub-km scales of the HRRRNEST domain.

3.3.4.4 Future Work and Recommendations

Now that a complete 1-year period of reforecast data is available in the WFIP2 Data Archive and Portal, other possibilities exist for train-test splits that may offer more robust results with larger in-season sample sizes available for the algorithm training periods. For example, the eight additional months could be set aside for training, while keeping the four existing months as out-of-sample test periods. Even better, a 12-fold cross-validation strategy could be used, rotating through the full data set one month at a time, yielding 11-month sample sizes for algorithm training periods. If longer training periods of about a year can be gathered, it should result in more skillful statistical models. In turn, this would yield a more accurate estimate of the potential industry benefits as a result of WFIP2 modeling improvements.

Since many of the envisioned modeling improvements in WFIP2 slated for the terra incognita scales did not get completed, it was not a surprise that the relative improvements available from the HRRRNEST simulations were not as substantial. Most of the modeling improvements implemented by NOAA stand to benefit the HRRR at operational resolutions above 1-km grid spacing. For the time being, because the primary benefits of WFIP2 will be realized through the operational HRRR model, it makes sense to more thoroughly examine the differences between the control and experimental WFIP2 HRRR at the 3-km operational resolution. In future years, after more gray zone modeling improvements are implemented into WRF, the impact on wind power forecasts at sub-km scales can be revisited.

3.4 Decision Support Tools

As lead of the WFIP2 Decision Support Tools (DST) team, Vaisala organized work to develop decision support algorithms for complex terrain phenomena observed in the field study. The motivation behind the DST team's work was to connect the WFIP2 project with the wind energy industry and to ascertain its relevance. The work was driven by two primary questions:

- How can we convey the possible impacts of complex terrain phenomena?
- Can we create actionable alerts that will improve situational awareness and reduce decision-making time?

To answer these questions, Vaisala organized and carried out a set of sub-tasks to design algorithms, build a prototype, and collect industry feedback. The unifying objective was to design a light-weight alerting prototype for complex terrain phenomena that drive wind power volatility, which was fully probabilistic and could be used to facilitate an interactive discussion with industry partners on the features and potential utility of the tool.

Coordination amongst representatives from WFIP2 project participants from NOAA, NREL, and PNNL at DST team meetings facilitated a robust exchange of ideas. After some iteration, the team reached consensus on the phenomena of primary interest and on the initial algorithm design. Cross communication with the Uncertainty Quantification (UQ) team established linkages to parallel efforts underway at PNNL, NREL, and TTU (a Vaisala team member) and yielded future plans to incorporate uncertainty information into the algorithm itself. Due to a shortened timeline for data analysis compared to the original WFIP project plan, Vaisala implemented the DST algorithm for only a single example complex terrain phenomenon. Team discussions laid the groundwork for design of additional phenomena, should future funding become available. For the single example type, a history of

probabilistic alerts was generated over the multi-season retrospective forecast period for WFIP2. These comprised four month-long periods (one calendar month from each season) during the field study. Creation of the alert history enabled for computation of overall validation statistics and an initial assessment of its performance. Vaisala modified its commercial interface to build a prototype tool for display of the alert information as a vehicle to collect industry feedback on its presentation format and potential usefulness.

3.4.1 Methodology

The process we followed began with identification of the complex terrain phenomena observed during the WFIP2 field study. This part of the effort drew on work from the larger WFIP2 team, including contributions from the Event Logging and Verification and Validation (V&V) teams. We relied heavily on the WFIP2 Event Log, a daily recording of the dominant observed weather conditions in the Columbia River Basin during the 18-month field campaign. The Event Log contains a subjectively applied weather taxonomy, identifying which of five major categories of meteorological phenomena were active during each day, the performance level of the operational HRRR model forecasts, and the relative importance of the event for the wind energy industry. From the Event Log, we were able to identify the most important event types that occurred frequently during the field campaign and were suspected as having the largest impacts on wind operators. These periods are also identified in the Common Case Study Set, where the selected sub-set of interesting case days are annotated with labels for significant power ramps (up or down) or periods of power volatility.

We also utilized the WFIP2 Phenomenon Identification and Ranking Table (PIRT), which was a tracking tool developed by the V&V team to collect the priorities and reasons behind the many proposed team-wide validation studies and to connect them with the available field instruments that would be needed to carry out the work. From the PIRT, we were able to identify the specific instruments likely to be the most helpful for identifying each type of event occurrence during the field campaign. The goal of the team was to automatically label event occurrence (as a binary class) using the field campaign data.

A principal objective from the WFIP2 science plan was that the DST algorithm should produce probabilistic guidance, and if possible, incorporate the work of the UQ team. Because of the sequencing of work during the WFIP2 project, the UQ team was only able to produce ensemble simulations for a limited set of case studies late in Budget Period 3. UQ simulations were completed for 16 cases before the project concluded, but this was an insufficient sample size to fully develop and test the candidate methods for integration with the DST algorithm. Instead, we stopped at a conceptual approach to the design and will leave it for future studies to carry forward and test the efficacy of these plans. Nevertheless, we did succeed in implementing a method that achieves the basic WFIP2 science goal of having a probabilistic alert. This was achieved through the use of modern machine learning methods, many of which enable the straightforward development of classifier models, which can produce either probabilistic or deterministic predictions of binary (or multi-category) events. Amongst many candidate techniques, ultimately we chose to implement random forest classifier models, as these achieved the best performance within our experimental period. It should be noted that the classifier models we implemented were only based on candidate features from the control and experimental HRRRNEST (Δx = 750 m) model runs and therefore don't incorporate initial condition or model parameter uncertainties. A future advance to the method would incorporate ensemble simulations, possibly produced via a formal UQ methodology, in order to better sample the flow-dependent uncertainties in the model simulations of relevant meteorological parameters. The candidate features from HRRRNEST simulations are used as inputs for training random forecast classifiers, which then produce probabilistic event-based forecasts on an independent test set.

3.4.1.1 General Procedure

The general procedure we developed for the DST algorithm follows below, accompanied by some notes explaining further details on each step:

i. Label times when complex terrain events of interest occurred.

Ideally this step is automatable, in that some rule set can be applied to observed data, and the result is a binary class label (0/1) for each time period. Otherwise, a brute force approach may be necessary requiring human effort and subjectivity to annotate each time point. This approach can naturally be extended to accommodate multi-class event types or real-valued indices over the [0,1] interval to add the notion of event severity or likelihood.

ii. Label times when power ramp events or significant power volatility occurred.

This step is also best, if automated. For identifying power ramp events automatically, we relied on the Ramp Tool & Metric (RT&M) developed during WFIP1 (Bianco et al., 2016) rather than inventing new techniques. Annotation of power volatility events would require new methods for automatic identification, and we did not end up pursuing these further in this study.

iii. Find the overlap times of these events from steps i and ii, creating merged labels for times with specific ramp-causing phenomena.

This step is straightforward if a logical operator on the two labeled time series can be used, like intersection. However, it could also be important to account for time lags or intermittency which could complicate this step and introduce additional configuration parameters. The merged labels become our target data time series, which we use for training predictive models for alerting on these joint conditions.

iv. Extract/engineer features from candidate predictor data sets.

This step involves pre-processing and data set manipulation. For NWP model output data, extraction includes some important details, such as the choice of spatial interpolation method (bi-linear from the four nearest grid points), time interpolation method, and whether additional derived variables not part of the standard model output (e.g., lapse rate) are calculated. Other input data types could also prove valuable depending on the target phenomenon and forecast horizon, such as recent observations from satellite, radar, nearby profilers, or regional surface observations.

v. Combine with uncertainty quantification information.

This step is optional, since ensemble NWP simulations would significantly multiply the available number of candidate features to be used in the DST algorithm making the following step difficult without a very long historical training data set. If feasible, we encourage the use of a combined initial condition and physics ensemble, since important uncertainties arise from both sources for complex terrain phenomena depending on the type and time of day Smith (2018). To reduce the multiplicative impact on the number of available NWP fields and contain the dimension of training data set, we envision limiting the feature set to only summary information from the ensemble using statistics like mean, median, minimum, maximum, standard deviation, or percentiles (as in Roulston et al. 2003). Another possible option is to use the Smith (2018) approach to down-select a sub-set of ensemble members which exhibit important characteristics that well span the range of uncertainties of interest (e.g., fast, medium, and slow members for a ramp event).

vi. Train classifier model.

This step involves training a statistical model, using merged event labels from step 3 as the target data and engineered NWP-based features from steps iv and v as the training data. A generalized linear model can be used, like for probability of precipitation forecasts using traditional Model Output Statistics approaches. Alternatively, there are many types of supervised learning techniques now widely available for classification problems like this one. Here, we want to choose methods that can produce probability forecasts of the class labels.

vii. Apply classifier model on test data, obtaining probabilistic class predictions.

This step involves out-of-sample execution of the trained model from step vi to produce probability forecasts for the class labels and to measure the model performance on previously unseen data.

viii. Apply a post-processing method to calibrate the forecast probabilities.

This final step is optional, depending on the raw model performance and the availability of sufficiently large training data sets. If the raw model produces un-reliable forecast probabilities, then a calibration step may be necessary to achieve good reliability, while attempting to retain the skill (resolution and sharpness) in the original model output.

3.4.1.2 Data

Actual power generation data recorded at the wind project substation level are the basis for the DST algorithm work. Vaisala team partners contributed this data directly to the project for 20 individual wind projects within the BPA control area. Some project partners contributed data at the wind turbine unit level, but this more detailed power and wind data was not used in the DST analysis. Additionally, the BPA fleet aggregate power generation was utilized, since it is freely available and published in near real-time on the BPA website (https://transmission.bpa.gov/Business/Operations/Wind/twndbspt.aspx). The DST algorithms were developed on the BPA fleet aggregate and on the individual site level, with a single wind project in the center of the WFIP2 study area near Wasco, Oregon called Klondike III. The objective was to have at least one example for each level of aggregate wind power forecasting to evaluate the method. All observed power data was converted to 1-hourly averages in this study. We did not attempt to analyze sub-hourly forecast skill.

The model forecast data for the DST work comes from the provisional WFIP2 HRRR reforecasts available from the Data Archive and Portal (https://a2e.energy.gove/about/dap). These simulations are "cold-start" configurations with initial conditions taken from the RAP and run twice per day (every 12 hours) through forecast hour 24. Otherwise, the WFIP2 HRRR (Δx = 3 km) utilizes the same domain setup and physics suite as its operational counterpart. A very high-resolution nested domain, nicknamed HRRRNEST, is run at Δx = 750 m inside the WFIP2 HRRR with a delayed start of 3 hours. Although the WFIP2 HRRR and HRRRNEST reforecast simulations were eventually created for one full year of the WFIP2 study period, at the time of this analysis, only four 1-month periods were available for download. These months were April, July, and October of 2016 and January of 2017, calendar months which are centered during each of the four seasons. We examined only the control and experimental HRRRNEST runs in this study due to time-limited resources in the data analysis phase of the project. The HRRRNEST runs target the terra incognita scales (200 m < Δx < 1000 m), where PBL parameterizations are known to struggle, but better resolve important terrain features of the region including the Columbia River Gorge. A more complete analysis of both WFIP2 HRRR and HRRRNEST simulations for the full year of available reforecast data is now possible, but was not ready by December 31, 2017 to be in time for this initial DST work.

3.4.1.3 Algorithm Example: Cold-Pool Mix-Out Up-Ramps

The WFIP2 research team placed particular importance on better understanding of meteorological processes influenced by complex terrain that lead to rapid changes in wind speed. One event type that remained in focus throughout the study was cold pools and their evolution. These events were predominantly cold season phenomena that were the second-most frequently observed event type during the measurement campaign (Wilczak et al. 2019). During the decay phase of these events, the sudden transition in the near-surface winds from light or easterly to strong westerly or south-westerly winds results in a strong upward ramp of wind speed through-out the Columbia River Basin. Wind power at the BPA fleet aggregate level is often marked by periods of low or near-zero production during these long-lived cold pool episodes punctuated by a rapid increase at the end (McCaffrey et al. 2019). The timing of the mix-out of the cold pools has been a notorious short-coming of mesoscale NWP models and was a point of particular emphasis for targeted physics improvements during WFIP2 (Olson et al. 2019). Because the WFIP2 research team spent significant effort studying this particular event type and modeling improvements were realized because of it, the DST team nominated this phenomenon as the

best candidate on which to begin analysis. We followed the general procedure outlined above and call out the steps specific to the analysis of cold-pool mix-out up-ramp events below:

i. Labeling cold pool events.

For identification of long-lived cold pool events in the Columbia River Basin during the field study, we relied on the technique described in McCaffrey et al. (2019). In this approach there are criteria for stability, wind, and duration. The stability criterion involves the calculation of the temperature lapse rate in the lowest 1 km above ground level and ensuring that it must be greater than -6.8 C/km. The layer average wind must either be easterly or below 3 m/s if it is outside the NE to SE sectors. After meeting the first two criteria, to distinguish the cold pool from diurnal radiatively driven events, its duration must last more than 20 hours with gaps no larger than 3 hours. In the McCaffrey et al. (2019) long-lived cold pool study, these criteria were used at a combination of observing locations within the WFIP2 study area. Cold pool event times were labelled only from conditions observed at the Wasco, OR location, where we made the assumption that profiling observations taken there were representative of the basin wide conditions. The time resolution of the event labels were limited to hourly.

ii. Labeling significant up ramp events.

Ramp events in wind power generation are identified through automated labelling of the actual time series from a wind project or from regionally aggregated observed values. As discussed at length in Bianco et al. (2016), there is no commonly accepted definition for a ramp event. No strict definition is available and the parameters can vary based on the needs of the end user. For this reason, a more flexible tool called the Ramp Tool and Metric (RT&M) was developed as part of the scope in the original WFIP study. The RT&M tool is publicly available online2 and provides flexibility to choose from three different methods of ramp event identification. For simplicity, we chose a single method of ramp event identification, the min-max method, which finds the maximum power amplitude change (Δp) within a sliding time window of length WL. This method avoids some problems with the fixed-time interval method and is simpler than the explicit derivative method. It can be argued also that it results in more intuitive ramp definitions which start and stop at the peaks and troughs of the power time series. Although the tool supports any reasonable choice for Δp and WL, to avoid running a large number of DST model simulations, each one with a unique version of labelled data owing to a different ramp definition, we chose a set of functional default values. We selected Δp as 15% of the installed capacity, which amounted to minimum power change of about 720 MW for the BPA aggregate and about 33 MW for Klondike III. We set WL as 12 hours for BPA and 6 hours for Klondike III. After applying these settings and using the RT&M to label the observed power time series, we retain only the positive events that are associated with upward directed changes in power output. The time resolution of the labeled data was limited to one hourly. If the actual ramp start or end time fell within an hour, then that whole hour was labeled as part of the event. The native resolution of the observed data was retained to define the start and end times. For example, the BPA aggregate power has time resolution down to the 5-minute level. While the choices were based on Vaisala industry experience, but also somewhat arbitrary, the ramp definitions settings are modifiable and were specifically discussed during the industry feedback sessions.

iii. Finding the times when cold-pool mix-out and up-ramp events overlap.

We marked the end time associated with each cold pool event labeled in step i calling that the official mix-out time. Then, from step ii, we retrieved the start and end time of each up-ramp event. If the cold pool end time intersected any up-ramp by falling in between the start and end times of that event, then we created a new labeled time series for that up-ramp event marked as true. Because the Wasco profilers only measure one part of the study area and mix-out times can vary throughout the basin depending on the scouring processes at work, we wanted to also allow for a time offset. This step seemed especially appropriate in dealing with the BPA fleet aggregate power, which is a summed reflection of the mix-out process happening in different regions at different times. The observed up-ramps in the BPA aggregate time series could be composed of several smaller contributing up-ramps as the

[2] http://www.esrl.noaa.gov/psd/products/ramp_tool

cold-pool is eroded at higher elevations or western-most wind projects first. Therefore, we chose to allow for a 1-hour tolerance between the end time of the cold pool and the start time of the up-ramp event. This threshold was chosen by inspection of the time series. An example of the merged labeling procedure for the BPA fleet aggregate power is shown in Figure 63.

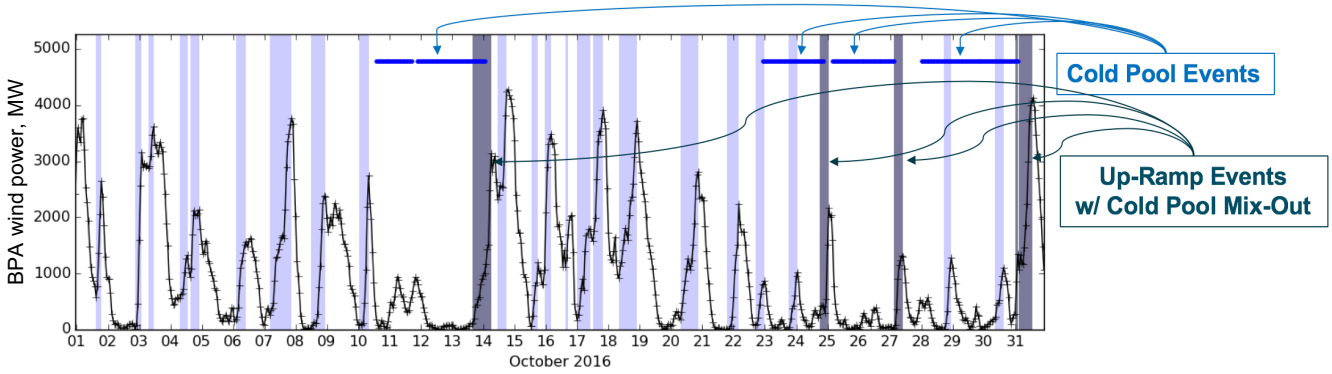

Figure 63. Observed hour-averaged power generation time series from the BPA fleet aggregate power during October 2016 (black). Labelled up-ramp event times are shaded light blue. Connected blue dots signify periods labeled as long-lived cold pools at Wasco, Oregon.

iv. Extracting and deriving important model input features.

For NWP candidate features, we used the control and experimental HRRRNEST model output bi-linearly interpolated to the Wasco airport location from the four surrounding grid points. No time interpolation was necessary as we used only hourly model output to match up with our hourly resolution labeled event data. We retained only forecast hours 4 through 15 from each model run that were 12 hours apart in order to stitch together a single complete training data set for all available valid times within the four one-month retrospective periods. A large variety of standard model output variables from HRRRNEST are available in the DAP and we utilized all of the 2-D surface fields as well as the 3-D fields at the native model vertical levels. To supplement the raw model output, we derived some additional variables that we hypothesized might be relevant for the cold-pool mix-out process. These additional model output variables were:

- The potential temperature vertical gradient between the surface and 300 m above ground level
- The potential temperature vertical gradient between the surface and 1.5 km above mean sea level
- The depth of the cold pool layer, as defined by an upward search for the maximum positive temperature lapse rate below 4 km
- The layer average wind speed within the cold pool layer
- The layer vector-average wind direction within the cold pool layer
- The wind speed at the model level immediately above the cold pool layer
- The u-component of the wind at the model level immediately above the cold pool layer

Although we restricted ourselves to focus solely on the usage of relevant NWP data here, other input data sets could have been used. For example, at short forecast horizons less than about 6 hours, observed or extrapolated data taken from observed conditions can often add significant predictive skill. For example, we could have added lagged observations from the wind profilers and thermodynamic radiometers located nearby within the study area to boost the predictive skill of the models at these shorter forecast horizons, but this potential improvement was not investigated here.

v. Incorporating UQ information (optional).

As mentioned earlier, we elected to skip this step and did not include data from UQ simulations for this example implementation of the DST algorithm. UQ simulations were only available for 16 days, meaning that we would have needed to invent a new method to synthesize ensemble forecast data for the remainder of the cases in the retrospective period in order to proceed.

vi. Training the statistical classifier models.

To compose a training data set, we assembled a history of target data using the merged labels from step iii and the NWP candidate features from the WFIP2 HRRRNEST simulations, as described in step iv. Statistical models were trained using standard classification methods readily available through the scikit-learn Python package, including support vector machines and random forest decision trees. Random forests produced the best results for the size and scale of the classification problem here, with a feature dimension of about 100 variables and a sample dimension of about 2500. To optimize the hyper-parameters of the classifiers, we employed a fixed grid search using 4-fold cross-validation.

vii. Testing and application of the model.

Due to limited sample size of the data in the study, with only a one-month sample from each season available, we could not afford to split the available periods into fixed training and testing sets. Instead we employed a k-fold cross-validation approach to assess how the results of our models would generalize to independent data. We divided the data set into eight folds, each one of approximately a half-month in length, leaving out one half-month partition as testing data and using the remaining 3.5 months as training data. We rotated through the folds, using each half-month period as a validation set. Days 1-15 were separated from days 16-end of the month. This had the benefit of including half a month of in-season data within the training data set when predicting the other half-month from the same season. With calendar month splits only using 4 folds, it would not have been possible to include in-season training data for the target periods with the available data set.

viii. Calibration of the forecast probabilities (optional).

We considered options for post-processing the probability forecasts to improve their reliability, but all available methods would have required an additional split of the data into training and testing sets to tune the parameters of the methods. Because we were already constrained by sample size, we did not want to further subdivide the limited data set available for model training. Furthermore, we observed that building classifier models for generic up ramp event target labels resulted in fairly reliable forecast probabilities that didn't require any post-processing. Therefore, we chose to skip this step and leave the cold pool mix-out up-ramp classifier forecasts in raw form.

3.4.1.4 Validation Strategy and Performance Metrics

To measure performance of the DST algorithms, we focused on a few standard validation tools for probability forecasts. To assess the agreement between the forecast probability and the mean observed frequency, we used the well-known reliability diagram (Wilks 2006). Reliability of the forecasts is estimated by the proximity of the plotted curve to the diagonal. We present reliability diagrams together with a histogram of forecast probabilities to convey the sharpness in the forecast values. As an overall summary of the skill of the classifier model forecasts, we computed the area under the relative operating characteristic curve (AUC) or ROC Area (Mason and Graham 2002). Implicit in the calculation of the AUC is that increasing probability thresholds are chosen to convert the forecast into a binary yes/no prediction and score it in the traditional categorical way. For each critical threshold, we can compute a corresponding 2x2 contingency table, and we can derive common summary metrics of the performance like the probability of detection (POD), false alarm ratio (FAR), and the equitable threat score (ETS). The AUC summarizes the potential usefulness of the probability forecasts over all of the possible decision-making thresholds.

Because our target labels are highly imbalanced, due to the infrequent nature of our example phenomenon, we must be cautious about using a metric like the AUC to discern between the skill of two different models. This is because the number of correct negatives can be very large and consequently a large change in the number of false alarms results in a relatively small change in the false alarm rate. Note that,

$$false\ alarm\ rate = \frac{false\ alarms}{false\ alarms + correct\ negatives}$$

with the correct negatives count appearing in the denominator. Therefore, AUC is not very sensitive to changes in model skill when the class imbalance is high. For this reason, we also compute the area under the precision-recall curve, which is a metric that is more sensitive to changes in the false alarm count. Precision, also known as the success ratio in the meteorological literature, is defined as:

$$precision\ (success\ ratio) = \frac{hits}{hits + false\ alarms}$$

and summarizes what fraction of the forecast yes events were correctly observed. It is the complement of the FAR. Recall is the same as POD. Both precision and recall do not include a term for the correct negatives, so naturally a precision-recall curve is much more sensitive to changes in the number of false alarms. The area under the precision-recall curve is also called the average precision score. This is because the area under the precision-recall curve is computed as the weighted average of the precision achieved at successively increasing probability thresholds. Each weight is defined as the increase in the probability of detection from the previous critical threshold.

To measure the decision-making impact of the probability forecasts from an economic impact perspective, we utilize the value score (Richardson 2000, Wilks 2001). The value score assumes a cost-loss decision model and is usually presented as a curve over all potential cost-loss ratios between zero and one. The value score conveys the percentage improvement in the expected expense associated with using the forecast information over time. It can be interpreted as the economic value between climatological (0%) and perfect (100%) information. The value score curve has a peak value at the cost-loss ratio that matches the observed relative frequency of the target event.

3.4.2 Results

Performance of the decision support algorithm outlined above is assessed by the set of diagnostic tools and accuracy metrics described in the methodology section above. Up-ramp events caused by mix-out of long-duration cold-pools in the Columbia River Basin are forecast retrospectively over the multi-season study period separately for the HRRRNEST control and experimental model simulations. The results presented here compare the performance of the control and experimental decision support models for every hour over 121 days during April, July, and October of 2016 and January of 2017. The total sample size for the comparisons is 2904 hours. Out of this total period, 811 hours or 27.9% of the times, are identified as part of up-ramping periods according to the criteria for the BPA fleet aggregate power. Only 70 of these up-ramping hours, or 2.4% of the total times, are tagged as being associated with cold-pool mix-out events. There are only 8 unique up-ramp events associated with cold-pool mix-out periods identified during these study. The up-ramping period for each event lasts long enough to collectively amount to a total duration of 70 hours. Because of the small number of identified event occurrences, this phenomenon can be considered rare and this makes accurate forecasting more difficult. Validation of the

forecasting performance for rare events is also difficult and the level of confidence associated with our results should be considered to be somewhat low.

Table 18. Overall performance of the probability forecasts for BPA fleet aggregate power up-ramp events associated with cold-pool mix-outs during the four-month, multi-season retrospective period during WFIP2. Forecast skill is measured by area under the ROC curve.

Cold-Pool Mix-Out Up-Ramps	HRRRNEST control Random Forest Classifier	HRRRNEST experimental Random Forest Classifier
AUC	0.723	0.763
Average Precision	0.056	0.063

In Table 18, the overall performance of the cold-pool mix-out up-ramp probability forecasts for the BPA fleet aggregate power are summarized by the area under the relative operating characteristic curve (AUC) and the average precision score. The AUC for the HRRRNEST control decision support model is 0.723 and for the HRRRNEST experimental model is 0.763. Average precision scores are 0.056 for the control model and 0.063 for the experimental one. These statistics show a very slim advantage to the experimental model, but the differences are quite small compared to the likely uncertainty in the scores due to low sample size. Therefore, we assume the differences are not statistically significant. The AUC level conveys that the probability forecasts have some limited skill, but that it is rather modest compared to the no-skill AUC of 0.5 and perfect skill level of one. The average precision scores better account for the event rarity than AUC does, and these scores show the very small incremental value over the minimum precision score of zero, compared to the perfect score of one.

The calibration of the probabilistic event forecasts are shown with the aid of the reliability diagrams in Figure 64. The right hand panels are simply zoomed in views of the left hand panels in order to focus in on the lower left portion of the plot due to the rarity of the observed event. The left hand panels show all values from 0 – 1, whereas the right hand panels extend only from 0 – 0.15. Because of the limited sample size and also the rarity of the event during the study period, we omitted plotting values where the bin population was less than 30. Focus is given only to the bins with a minimum reasonable sample size, while bins with very low confidence statistics are avoided. The largest forecast bin exceeding the minimum sample size has a mean predicted value close to 30%, which is over 10 times larger than the observed frequency of the event. So, we can conclude that despite our limited range of plotted values, the reliability diagrams do span a reasonable range of forecast probabilities from well below to well above the normal likelihood of the event occurrence. Within the six lowest bins (forecast values ranging from about 0% – 10%) the experimental HRRRNEST classifier forecasts are much closer to the 1:1 line than the control in four of the cases. In the lowest bin, the two forecasts have nearly equal observed relative frequencies. In the two largest bins (forecast values ranging from about 15% – 30%), the observed relative frequencies for both forecasts appear to be much lower than the forecast probabilities for those bins leaving the calibration curves far from the 1:1 line in this part of the diagram. Overall, it appears that the experimental HRRRNEST classifier forecasts are have better reliability than the control, but only in the 0 – 10% forecast value range. Above that level, both forecast types appear to be unreliable and over-forecast the event occurrence.

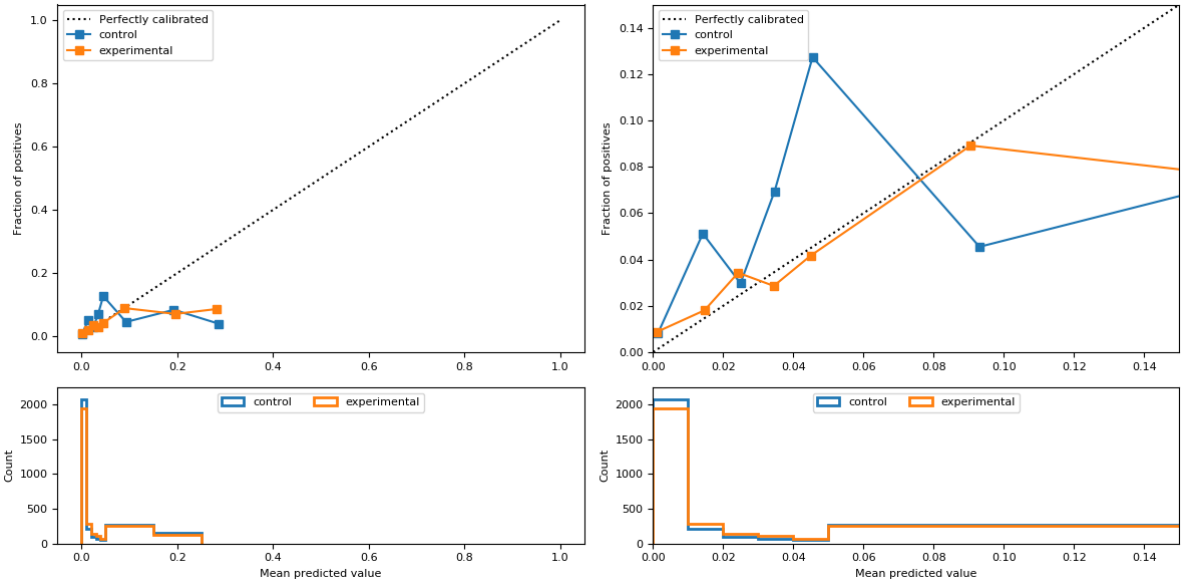

Figure 64. Reliability diagrams (top panels) for the control and experimental HRRRNEST classifier model forecasts of the BPA fleet aggregated power up-ramp events associated with cold-pool mix-outs. The corresponding histograms (bottom panels) show usage frequency.

To optimize decisions, a risk-neutral user with a sufficiently large budget should choose a critical threshold that is equal to the event likelihood. For our cold-pool mix-out example, such a decision-maker ought to choose the 2.4% threshold for converting the up-ramp probability forecasts to categorical (yes/no) ones. At this critical threshold, we can compare the resulting 2x2 contingency tables that are populated over the 2904 hours in our historical re-forecast period using the simple binary decision process. In

Table 19, we show the 2x2 contingency table for the control HRRRNEST classifier model at the 2.4% threshold along with the POD, FAR, ETS, and value score as summary metrics. In Table 20, we show the show the same statistics for the experimental model for comparison. At this critical decision threshold, the control forecasts perform a little better in all metrics. There are a larger number of event hits (53 to 50), a lower count of false alarms (847 to 896), and a smaller amount of missed events (17 to 20). Both forecast models have large FARs well above 0.9, with PODs only in the 0.7 – 0.75 range. The result is a positive ETS for both models, but only marginally above zero, since false alarms are penalized equally to misses in this metric. The value score takes into account the differing economic penalties between these outcomes by assuming that a false alarm only costs 2.4% of the amount of a missed event, since that is the chosen decision threshold. By this measure of economic value, both forecast models score well, getting about 40% of the way from a climatological baseline level toward having perfect information. The control forecast has a value score of 0.458 compared to the experimental value score of 0.398.

Table 19. 2x2 contingency table for control HRRRNEST classifier model forecasts at the 2.4% critical threshold with four summary statistical measures of the decision-making performance.

HRRRNEST control Random Forest Classifier Threshold = 2.4%		OBSERVED	
		YES	NO
FORECAST	YES	53	847
	NO	17	1987
POD		0.757	
FAR		0.941	
ETS		0.035	
Value Score		0.458	

Table 20. As in Table 19, except for experimental HRRRNEST classifier model forecasts.

HRRRNEST experimental Random Forest Classifier Threshold = 2.4%		OBSERVED	
		YES	NO
FORECAST	YES	50	896
	NO	20	1938
POD		0.714	
FAR		0.947	
ETS		0.029	
Value Score		0.398	

In Figure 65, we display value score curves as the cost of a false alarm changes and the critical threshold goes along with it for the cold-pool mix-out event example. The value scores are plotted as a function of the user's cost-loss ratio from zero to one. The value score is maximized at a cost-loss ratio that is equal to the relative frequency of the event, or about 2.4%, with values as seen Table 19 and Table 20 above. The value score curve drops off quickly to the left of the peak as the cost-loss ratio declines toward zero. The value score also decreases to the right, but less rapidly, as cost-loss ratios increase toward 10%. The experimental HRRRNEST classifier forecasts exhibit positive values scores over a broader range of cost-loss ratios than the control, despite the lower peak value. For cost-loss ratios in the range of 3% – 7%, the experimental forecasts perform better. Outside of the narrow range in cost-loss ratios from 0.5% – 7.5%, neither forecast shows an economic benefit to the decision maker since the value scores are less than or equal to zero in these zones.

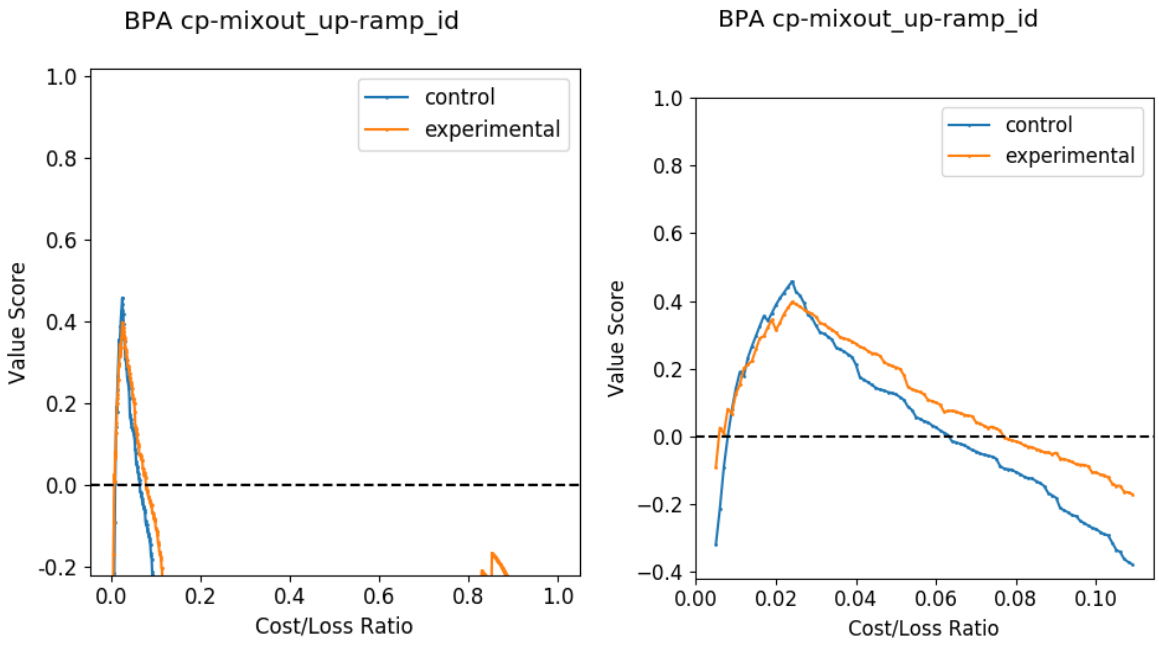

Figure 65. Value score curves as a function of the cost-loss ratio for the control and experimental HRRRNEST classifier model forecasts of the BPA fleet aggregated power up-ramp events associated with cold-pool mix-outs. The right hand panel shows the same data as the left hand panel, but with a zoomed in view to the cost-loss ratio range between 0 and 0.1.

For gaining context on these results, we also sought to understand how our example event forecasts perform relative to something more common. For this purpose, we compared the performance of classifier models trained separately for all up-ramp and down-ramp events in the BPA fleet aggregate power regardless of the cause. Over ten times the number of true events were available, with 811 hours (27.9%) identified as up-ramping periods and 952 hours (32.8%) labelled as down-ramping periods. In Table 21 we report the overall performance statistics for BPA up-ramps for both control and experimental HRRRNEST classifier models. The same results for the down-ramp follow in Table 22. There is no significant difference between the control and experimental model performance for either generic ramp event type. The skill for down-ramp events appears to be better than for up-ramps, with larger AUC and average precision scores across the board. For up-ramp events alone, compared to the relatively rare cold-pool mix-outs, the average precision scores are significantly higher, reaching above 0.5 in both cases. However, there is not an impact on AUC scores.

Table 21. Overall performance of the probability forecasts for BPA fleet aggregate power up-ramp events. Forecast skill is measured by area under the ROC curve (AUC) and average precision scores for the control and experimental HRRRNEST classifier models.

ALL Up-Ramps	HRRRNEST control Random Forest Classifier	HRRRNEST experimental Random Forest Classifier
AUC	0.719	0.727
Average Precision	0.501	0.513

Table 22. As in Table 21, except for BPA fleet aggregate power down-ramp events.

ALL Down-Ramps	HRRRNEST control Random Forest Classifier	HRRRNEST experimental Random Forest Classifier
AUC	0.804	0.804
Average Precision	0.659	0.662

In Figure 66, we show the reliability diagrams for the all up-ramp event classifier models (left) and the down-ramp event models (right). Overall, the forecasts show reasonable calibration and little difference between control and experimental versions. The up-ramp event forecasts seem to have less resolution in the middle forecast probability ranges between about 40% – 65% where the calibration curve deviates significantly from the 1:1 line. However, at low and high values, both forecast types perform well. Because both types of generic events are more common, forecast bins are well populated throughout the whole range.

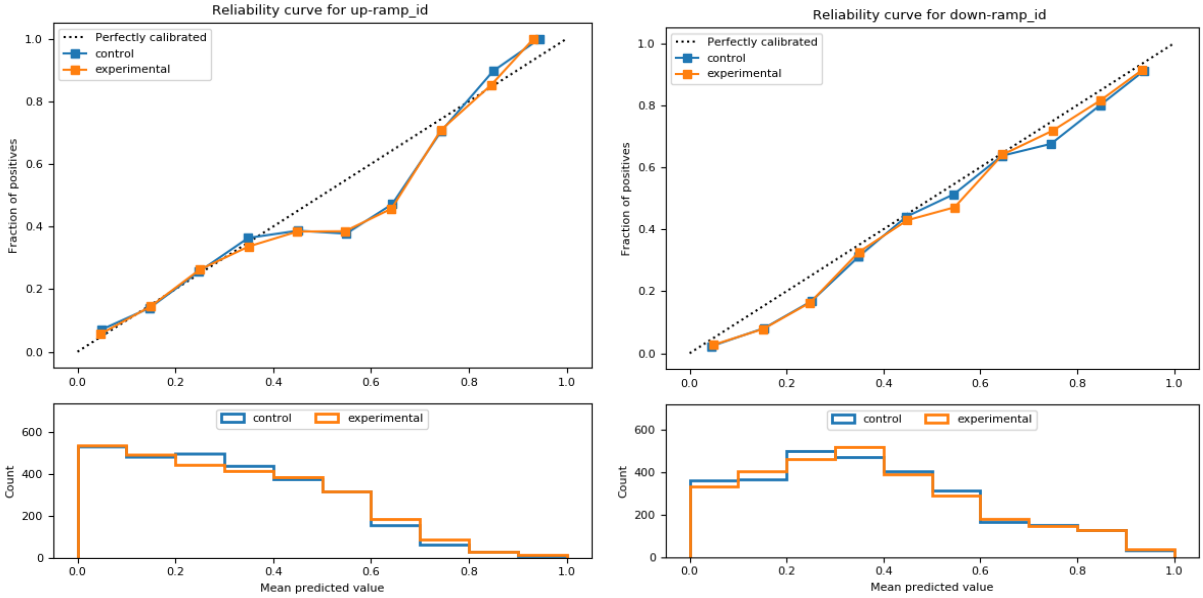

Figure 66. Reliability diagrams (top panels) for the control and experimental HRRRNEST classifier model forecasts of the BPA fleet aggregated power up-ramp events (left panels) and down-ramp events (right panels). The corresponding histograms (bottom panels) show usage frequency of each forecast value range.

The value score curves shown in Figure 67, further verify that there is no appreciable difference in skill between the control and experimental models for generic up and down ramp event forecasting. Compared to the value score curves for the cold-pool mix-out events, a broader range of positive value scores is evident, but the peak value scores are similar. For down-ramp events, the positive value score curve extends over nearly the entire range of cost-loss ratios. For up-ramp events, the positive value score curves are interrupted in the area of low resolution between about 40% – 65% that was also evident in the corresponding reliability diagram above. Otherwise, the up-ramp value score curves are positive from about 5% – 40% and again from about 65% – 100%. The peak value scores occur at much

larger cost-loss ratios than for the cold-pool mix-out up-ramp forecasts, as we should expect, where they are equal to the event frequency of occurrence (27.9% for up-ramps and 32.8% for down-ramps). The peak value scores range from as low as 0.319 for experimental up-ramp event forecasts to 0.456 for control forecasts of down-ramp events.

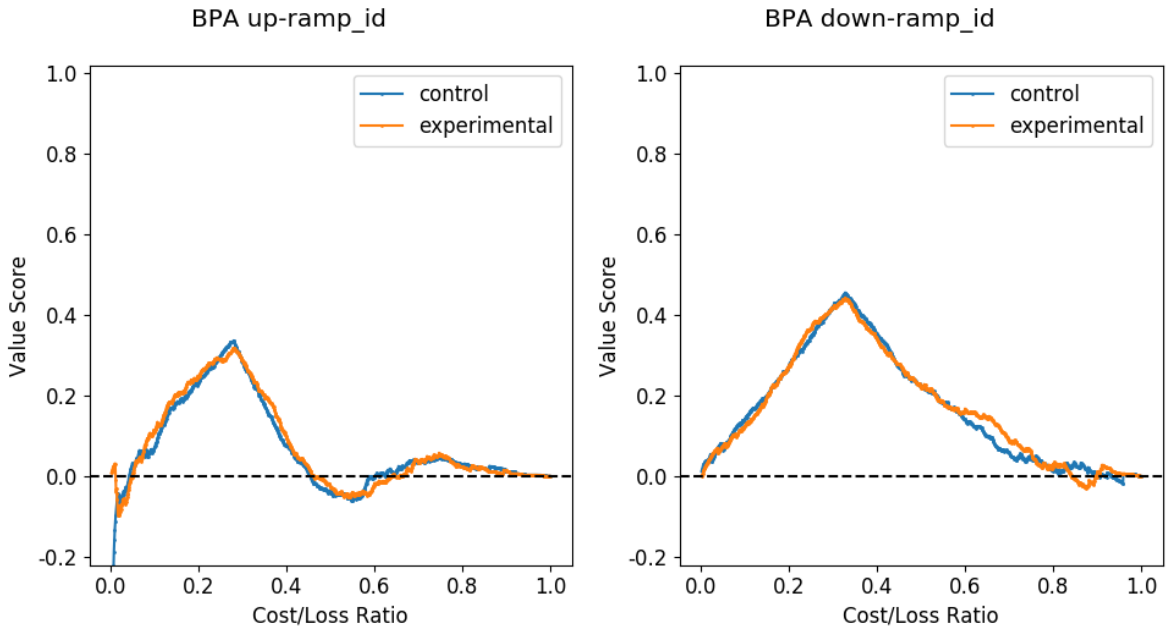

Figure 67. Value score curves as a function of the cost-loss ratio for the control and experimental HRRRNEST classifier model forecasts of the BPA fleet aggregated power up-ramp events (left panel) and down-ramp events (right panel).

At the individual wind project level, we are unable to test whether the cold-pool mix-out event forecast performance differs from the fleet level aggregate. After implementing the labelling procedure at a representative Wasco area wind project, called Klondike III, only 3 hours were identified as true events. With only 3 positive samples in the dataset, we are unable to train a robust classifier model to make out-of-sample forecasts. In fact, the 3 hours with positive labels are part of a single up-ramp event on January 19, 2017, meaning that when forecasting for the second half of January 2017 in our cross-validation procedure, there are no observed events within the available training data set from the other re-forecast periods. This exceedingly low sample size is due to a few factors, including at least: (1) the shorter ramp event duration at the individual wind project level, (2) localized effects near the project scale that mask the primary up-ramp event associated with cold pool mix-out, or (3) a time offset between the marked end time of the cold pool and the start time of the up-ramp event that is outside the chosen tolerance we defined in the original procedure. The first two effects are not easily addressed, so we chose only to pursue item (3) above by varying the time tolerance used to define intersection of the two event labels. Increasing the time tolerance first to 3 hours, then to 6 hours, did not increase the number of intersecting labelled hours at Klondike III. Only after increasing the time tolerance to 9 hours, did we pick up 4 new events from October 2016, as seen from the time series plot in Figure 68. However, this 9-hour time tolerance setting only added an additional 13 hours, yielding a sum of 16 total hours for which true intersecting events were identified. Upon inspection of Figure 68, and for similar time series plots corresponding to each of the other four month-long re-forecast periods, a 9-hour time tolerance seemed more than generous as the maximum amount. Increasing the time tolerance to 12 hours or more seemed physically unrealistic, since the cold pool labelling algorithm used the Wasco Airport profilers located quite close in proximity to the Klondike III wind project. We did not investigate further increases in the time tolerance threshold, since it seemed likely that any increase in sample size would be gained through

incorrect labelling of up-ramp events caused by a phenomenon other than the mix-out of the local-area cold pool.

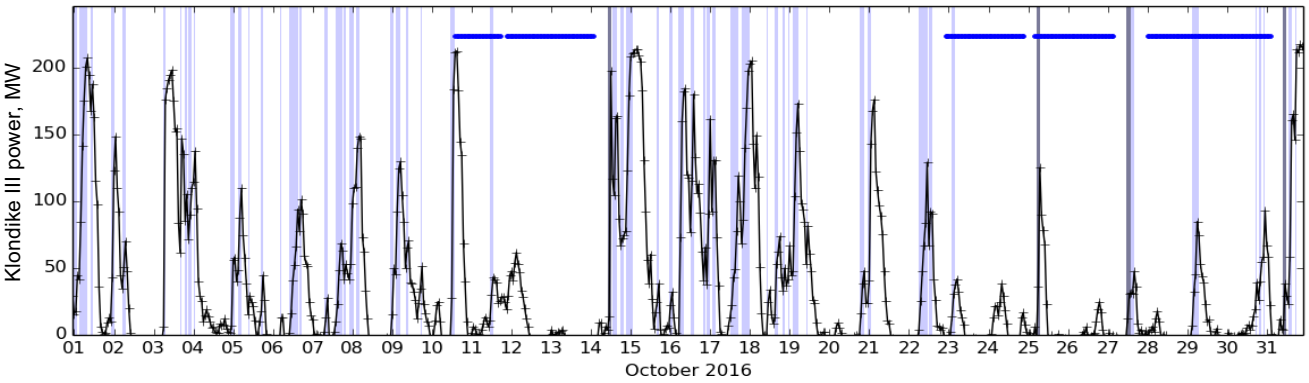

Figure 68. Observed hour-averaged power generation time series from Klondike III wind project near Wasco, Oregon during October 2016 (black) with labelled up ramp and cold pool event periods as in Figure 63. The intersection of the end times for cold pool events with the start times of up-ramp events, with a 9-hour time tolerance, are shaded a medium blue. There are four events flagged encompassing 13 hours during the month.

For generic up-ramp and down-ramp events, which occur fairly often, we are able to confirm that performance at the individual project level is similar to the fleet aggregate. The frequency of occurrence of the events at Klondike III were at least as common as for the BPA fleet, despite the shorter duration of each individual event, with up-ramp events labelled on about 33% of the hours and down-ramp events identified at around 37% of the hours. Overall skill of forecasts at the Klondike III site was measured by AUC at around 0.7 and yielded average precision scores near 0.55 for up-ramp events and about 0.6 for down-ramp events. Value scores peaked around 0.34 for up-ramp events and close to 0.29 for down-ramp events. The reliability diagrams showed that the forecasts were close to perfectly calibrated, except for the mid-range up-ramp probability forecasts, just as observed for the BPA fleet level.

3.4.3 Prototype Tool

The prototype tool development work builds upon an existing user interface already available for paying subscribers to Vaisala's operational forecasting services through its Energy Dashboard. The existing tool, called the "Multi-Forecast Tool", is a configurable, adjustable graphical tool embedded in the Energy Dashboard interface, which is capable of displaying either real-time or historical forecast data from multiple data sources alongside observations. For wind power forecasting customers, a typical default view might include the real-time power forecasts for the facility or region of interest from one or two sources over next 7 days with recent observations plotted for the past 1 day. The Multi-Forecast Tool also has a feature called "Historic Mode", which allows the user to select any historical day and time during the subscription history to view old cases alongside the observed conditions that actually occurred. Because the available data for WFIP2 was all retrospectively generated, we relied on the Historic Mode in the Multi-Forecast Tool as a framework for displaying the experimental forecasts and facilitating conversations with our industry partners. This naturally brings the inherent advantage of hindsight to all parties, since the actual observations are available for viewing. Therefore, it does not replicate the real decision-making paradigm encountered in real-time when only forecast data would have been available.

Since a key requirement of the alert design in WFIP2 is that it is fully probabilistic, Vaisala focused on expanding the Multi-Forecast Tool functionality to accommodate a few new features in the initial prototype work. The starting point was a limited alert display, initially only available for deterministic forecasts of high wind speed conditions. When activated, the alert display provides shading over the

forecast time series for affected hours that exceed the selected high wind speed threshold criteria. The configuration of this high wind speed threshold is available through a settings menu that is accessed via a drop-down function in the upper right corner of the tool. We retained that general paradigm for the new probabilistic alert display and settings configuration, but expanded the types and styles available to a user. New alert types were added for generic up ramps and down ramps. For the prototype tool, we elected to proceed without the special ramp event type for the cold-pool mix-out phenomenon because of the event rarity and the limited skill obtained from the modeling work. However, the overall concept wouldn't change the presentation if we had added it as another option. Configuration options were added to the settings menu to allow the user to select the color of the shading for each ramp alert type, as well as the critical threshold in integer percentage points [0, 100]. A small icon indicating an active up-ramp alert (up arrow) or down-ramp alert (down arrow) represented in the same color as chosen for the shading was also added into the legend below the time series plot as a feature to help users quickly identify if attention should be paid to the forecast for that site. A screenshot of the Multi-Forecast Tool with these new prototype features and set to Historic Mode for an April 9, 2016 05:00AM PDT forecast of the BPA wind power aggregate is shown in Figure 69.

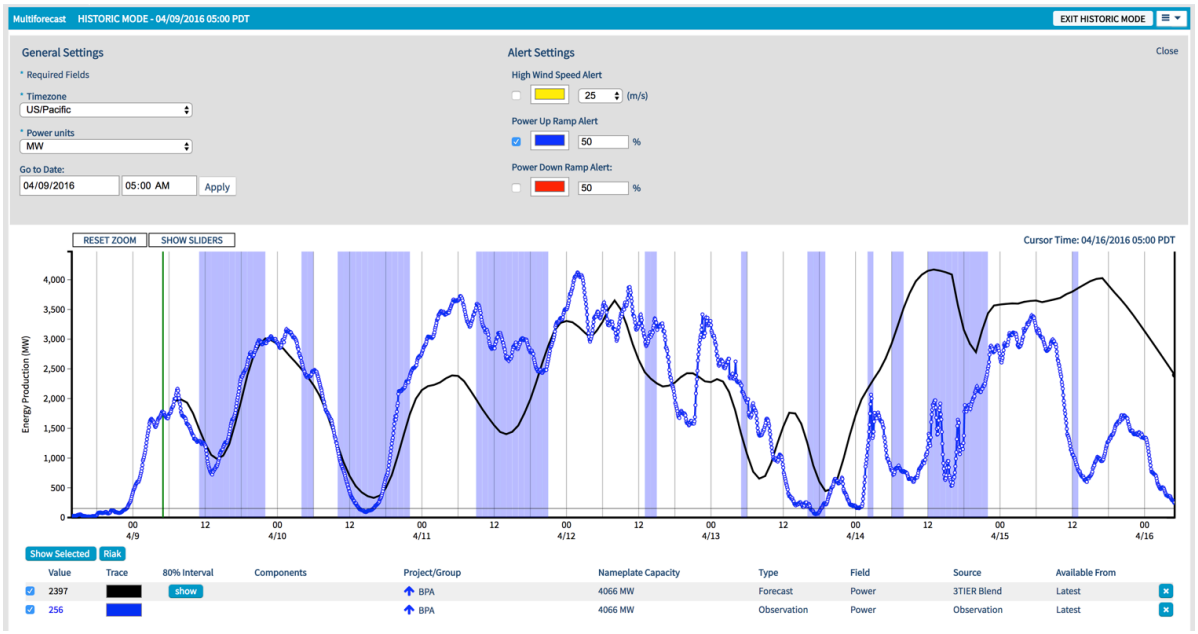

Figure 69. Screenshot of the prototype DST utilizing Vaisala's existing Multi-Forecast Tool with new alert types enabled. The tool is set to Historic Mode displaying the latest hourly-resolution forecast from April 9, 2016 at 05:00 PDT for the BPA fleet aggregate wind power (black curve) along with the corresponding 5-minutely power observations (blue dots). Only alerts for generic up-ramp events are activated and shown in the forecast time-series as blue color-shaded regions where the probability exceeds the 50% critical threshold. A corresponding blue up-arrow icon is shown in the legend for the affected project (BPA).

In the top menu above the main time series display, called "Settings", only the up-ramp alert type is checked as active, while the down-ramp and high wind-speed alerts are not checked and are therefore inactive in this example. Active alerts produce color-shaded areas on the forecast time series plot, but only for hours where the alert probability exceeds the critical threshold. The critical threshold values are configurable. For this example, the user has set the threshold at 50%, so only times where the up-ramp probability exceeds 50% are shaded blue. The color choice itself is configurable to ease cases of conflict with other active alert types, pre-existing time series of similar color, or color blindness of the user. Because the up-ramp alert is active and there are instances in the forecast that exceed the critical threshold, the up-arrow icon is shown in the legend for the BPA project.

In Figure 70, an example of the prototype DST is shown with the down-ramp alert type activated at the 50% threshold for the same historic forecast on April 9, 2016 at 05:00 PDT. Our example user has chosen a red color to delineate down-ramps from the blue shading used for the up-ramp alerts. The active alert icons are now shown in the legend as red down-arrows.

Figure 70. As in Figure 69, except with only generic down-ramp event alerts activated. The alerted hours are shown in the forecast time-series as red color-shaded regions where the probability exceeds the 50% critical threshold. A corresponding red down-arrow icon is shown in the legend for the affected project (BPA).

3.4.4 Industry Feedback

Our approach to collect wind energy industry feedback was to conduct interactive outreach meetings with WFIP2 project data partners. Two-hour sessions were set up with two project partners: Portland General Electric (PGE) and Avangrid Renewables. The general format of these meetings was organized to cover: (1) background information on the WFIP2 project, (2) the DST objectives and experimental results, (3) a live demonstration of the prototype DST for cases of expressed interest, and (4) discussion time using a pre-prepared questionnaire. The methodology and results of the DST work were shared through a set of presentation slides. Then, we transitioned to a web browser to display the Vaisala Energy Dashboard and the Multi-Forecast tool set to Historic Mode for the April 9, 2018 case, as shown in the figures above created from screen captures. This particular period was chosen as a good default example because it marked a particularly volatile week during the spring reforecast period that also figured prominently in the common case study set.

Using the April 9-16, 2018 period as the default view, the new features for display of the generalized ramp alerts were demonstrated along with the ability to change settings, toggle the alert types, and switch to other dates on request. The live sessions made for a more engaging discussion than would otherwise have been possible with only still, screen captures of the prototype tool. Nonetheless, it did not achieve the full live experience a user might get in real-time at the forecast desk when using the tool on a potential event that is yet to happen. Instead, historic events from the WFIP2 field campaign period that fell within the four 1-month retrospective periods used for the DST experiments were discussed. While this naturally limited the cases we could inspect together, it also helped focus the discussion to particular events of importance that we asked PGE and Avangrid to identify in advance of the sessions.

The survey questions posed to our project partners were the following:

- Do you have a troublesome past event in mind that took place during one of the 4 month reforecast periods we analyzed in WFIP2 (Apr, Jul, or Oct 2016; Jan 2017) that we can take a look at as an example to focus our conversation? (If not, use pre-chosen April 9-16, 2016 period as a default.)
- We have designed prototype alerts for generic up/down ramps, and one special alert for an up-ramp caused by cold-pool mix-out events observed during WFIP2 by the team. The guidance is probabilistic, so you can dynamically set a threshold to customize when the alert is active, given your personal (or business case) tolerance for false alarms.
- What do you think about the ability to customize the probability threshold to tune the alert for your needs?
- What do you think about the ability to change colors and toggle the alerts on/off?
- What do you think about the special alert? Is it useful to know what type of weather phenomena might be causing a ramp while being alerted about it?
- The prototype ramp alerts have assumed a minimum ramp size in their very definition (e.g., 700 MW for BPA, around 15% of installed capacity) and do not allow you to customize the ramp size and therefore prevent changing the alert probability.
- Is this a key limitation?
- If you need to alert on ramps of different/various sizes, do you need it to be fully customizable? Or can you live with a small, pre-selected list of ramp sizes that are important thresholds for your operations?
- Can customizable alerts help your scheduling/trading/operations decisions? Or is this heading in the wrong direction?
- What did we miss that is absolutely key for impacting your decision-making process?

As follow-up steps to the interactive sessions, the questionnaire was sent to our project partners in a document annotated with our notes from their verbal input to each item. They were invited to verify the existing notes and to provide any additional commentary or ideas. Screenshots of the prototype tool were sent alongside the document to share with colleagues in case other people in their organizations were interested and wanted to provide feedback. This additional input was collected and the overall findings were compiled and summarized in Table 23.

Table 23. Summary of wind industry feedback on the DST prototype organized by attribute.

Prototype DST Attribute	Industry Feedback Highlights
Customized Probability Threshold for Tuning Ramp Alerts	• Ability to change the critical threshold is vital • Best used by balancing authority for reducing reserve requirements during low risk times • Capability to turn alerts on/off is an important feature
Special Ramp Alerts by Phenomena Type	• Helpful for on-staff meteorologists, “pro” users • Probably information overload for real-time trader or balancing authority operator • Industry needs to figure out how to value this
Ramp Size/Duration Definition	• Options are needed; pre-set thresholds (aligned with balancing authority requirements) are preferred • When/where (timing and level) of the end of down ramp events is more important • Shorter ramp window is essential (focus in BPA is almost entirely on 1-hour ahead, not next 6 hours)
Potential Impact on Decision Making	• A useful education tool if both balancing authority and owner/operators have the same view • Could target tool usage toward improved negotiation for reserve capacity • Potentially useful tool if reserve costs could be dynamically input to help define actions
General Display	• Changing colors is nice to have, but not essential • Ability to display the event forecast probability itself as its own line on the time series plot would be useful

The overall impressions of the DST prototype were positive and there was general enthusiasm for such an interface. The tool was deemed to be of particular use for managing reserve capacity requirements, especially if the electric system balancing authorities and wind facility owner/operators can have the same view. To achieve this joint state of understanding, it was suggested that it could be approached from an educational standpoint at first, while experience was built by both sides using the tool operationally. At present, both parties oftentimes get their wind forecast information from different providers and the component models and tool sets can vary. One vision of the future is that when the risk of a down ramp event was low enough and both parties agree this is the case, the balancing authority could relax the standard amount of required reserves during this period.

Customization of the critical thresholds were seen as essential because that is how the ramp event risk and reserve sizing could be defined. There was a general recognition that different probability thresholds would be needed and choosing a fixed level would be too restrictive because reserve costs change. One suggestion was to consider making the tool even more dynamic by allowing it to accept these costs as an input data set. Other customization elements like color choice were considered nice-to-have features. Weather-regime specific information, like for special ramp event types, was likely to be considered information overload for the typical real-time trader, but possibly useful for any supporting meteorology staff or other “pro” users.

There was considerable input about the choice of the ramp definition used for the experiments and the DST prototype. While the default choices were based on Vaisala industry experience, but also somewhat arbitrary, it was anticipated that the ramp definition settings would need to be modifiable. The consensus opinion from the feedback sessions was that a pre-set list of options for ramp size and duration were

needed, and that the preferred approach to choosing the set would be to make them conform with the requirements of the balancing authority. Therefore, regional customization of the lists would be required and they may also need to be adjusted as the installed wind capacity of each system grows. Specific input was given about the need in the BPA region to focus on the next 1-hour period, although it was recognized that wholesale electricity scheduling timelines can also vary for other electricity system control areas. One particular area of identified need was for increased focus on the exact magnitude and timing of down ramp events, rather than just the overall risk of any significant event that qualified above some minimum thresholds. Anticipating the end time and level of down ramp events is a critical need, even though for many meteorological phenomena forcing these events, the predictability of such details may be quite low.

3.4.5 Summary and Discussion

A general algorithm was designed to frame complex-terrain phenomena studied during WFIP2 in terms of their potential for causing wind power ramp events. This decision support algorithm was tested on a limited portion of the WFIP2 data set and implemented for a single event type: up-ramps caused by the erosion of long-lived cold pools. The initial results show that probabilistic forecasts for these special events have limited skill. The average precision score is only about 5-6%, while for generic up-ramp and down-ramp events it can reach moderate values around ten times larger (50-60%). At the optimal decision-making threshold, we observe low equitable threat scores near 3% because even though probabilities of detection are above 70%, the false alarm ratios are even larger (above 90%). Only within a narrow range of critical thresholds (the 0.5-7.5% interval) can the forecast be used to economic advantage, where value scores reach up to a level of around 0.4, but drop off quickly on either side of this peak. This means that for cost-loss ratios less than 0.5%, a decision maker should always choose to protect against the event happening, despite its rare nature. For cost-loss ratios greater than 7.5%, a decision maker should never choose to protect against the event happening since losses aren't too expensive compared to the cost of protection. Because of the relatively infrequent nature of the DST example event type and the small sample size of observed events during the four-month reforecast period available in this study, it's likely these results are too pessimistic and can be improved with the use of the full year-long historical WFIP2 reforecast data set that is now available.

Although significant improvements were implemented and measured for raw model forecasts from HRRR during WFIP2, we were not able to detect any significant differences between the control and experimental runs as part of the DST work. One reason for this is that we only validated forecasts based on the higher resolution HRRRNEST, which showed smaller improvements relative to the control forecast than the lower resolution HRRR in atmospheric parameter evaluations (Olson et al. 2019). An additional reason is that after post-processing is applied, we should expect only smaller differences to emerge, since the statistical models act to correct inherent biases of each model. In our DST experiments for cold-pool mix-out up ramps, both classifier models are unreliable above the 10% forecast value range, substantially over-forecasting the event occurrence. The experimental classifier model seems to have better reliability in the 0-10% forecast value range and a larger interval of positive value scores. AUC and average precision scores are only slightly improved in the experimental runs. Conversely, the control classifier model exhibits slightly larger equitable threat scores and value scores at the optimal critical decision threshold.

We conducted the DST work with fixed ramp definition settings to retain simplicity and keep the number of experiments to a manageable size. However, it was known in advance, and confirmed through the industry feedback sessions, that any such models should be built for a range of event threshold definitions. Our industry partners preferred that these settings be selected based on balancing authority requirements. It follows that a full exploration of the utility of ramp event alert skill, whether for generic or special complex-terrain driven event types, ought to include results over a range of pre-defined ramp event magnitudes and durations.

A prototype DST was developed as an extension of Vaisala's commercial user interface for power forecast subscribers. We elected to include only generic up and down ramp alert types directly in the example tool, because these alerts were demonstrated to have a useful level of skill over a broad range of decision thresholds and occurred so frequently that it would be easier to facilitate discussion with our industry partners. Although we omitted the special complex terrain alerts from our worked example from the prototype, we discussed options for how such features could be included as part of the industry feedback sessions. The tool features are easily extended to include such special alerts as additional event types with their own individual check-boxes to be activated. Another suggested option was to have them appear as tool-tips or special alert messages noting the ramp event cause when triggered by their likelihood exceeding a critical threshold probability. Since our initial results demonstrated only low skill levels for these special alerts, our industry partners agreed that they were not yet compelling enough for operational use. If the DST methods described in this study can be further developed through more analysis and availability of longer training data sets to achieve higher levels of skill, it is possible that advanced users might see benefit, especially those with meteorological support teams. However, it was also clear from the feedback that the industry has no clear way to value this kind of contextual information, so the investment payoff is harder to justify.

3.4.6 Future Work and Recommendations

Industry feedback on the prototype DST was positive enough to justify continued work, while respecting public-private roles. The level of delivery customization and need for operational customer support is best suited for private sector to lead. The scope of consulting work likely includes establishing the set of supported ramp definitions, defined event types, and user interface features as well as maintaining updates to them. Public sector work focused on estimating decision-making value of the underlying information is essential to unblock industry adoption. Since private sector investment has such an uncertain payoff in these features, it is higher risk for industry to pursue them. Government and academic researchers could collaborate to extend this study to larger sample sizes and more ramp causing phenomena in order to better establish the practical predictability of different event types. For example, the DST team began work on a second example event type that was left incomplete: down ramps associated with gap flow decay. The wind industry has expressed a strong desire to improve predictive accuracy during down ramp events, with the goal to better anticipate their duration and ending level, so work in this direction could be of high value. Within these predictability studies of different event types, care should be taken to distinguish the skill of raw NWP model output and statistically corrected guidance since we know the relative improvement available is smaller in the latter.

This research report documents a general methodology that DOE labs could utilize to complete further work on DST algorithm design, if consistent with operating plans. For example, DOE effort in uncertainty quantification is ongoing and we have left that element unimplemented in this initial study. Since complex terrain phenomena have such a strong dependence on initial condition uncertainty, the use of single deterministic NWP runs to make probabilistic forecasts of these events is likely sub-optimal and leaves a promising avenue of improvement available.

Now that a complete 1-year period of reforecast data is available in the WFIP2 Data Archive and Portal, other possibilities exist for train-test splits that may offer more robust results with larger in-season sample sizes available for the algorithm training periods. For example, the eight additional months could be set aside for training, while keeping the four existing months as out-of-sample test periods. Even better, a 12-fold cross-validation strategy could be used, rotating through the full data set one month at a time, yielding 11-month sample sizes for algorithm training periods.

Finally, DOE could consider making available further applied research or human factors funding to pursue an extension of the DST time-series prototype into the spatial domain. Due to the shortened work time available in the final phase of the WFIP2 project, we weren't able to create a prototype that might be better geared for situational awareness over a regional control area where multiple wind projects are simultaneously operating and might be exposed to various complex terrain phenomena at different times.

A thematic, spatial map of complex terrain alerts could be designed and tested with a group of interested industry users.

4 Summary

WFIP2 was a multi-year, multi-institution effort to improve wind power forecasts in complex terrain. It included an 18-month observational field study and extensive development and analysis of parameterizations for near-surface physics in numerical weather prediction. The primary outcomes of the project are a publicly available observational data set, an advancement of understanding of the weather phenomena that affect hub-height winds in complex terrain, improvements to NOAA's HRRR model, and algorithms for decision support tools for wind power forecasts.

The WFIP2 observational data set is a comprehensive meteorological resource for a variety of atmospheric phenomena in areas of complex topography that enabled both the science within this project and will likely have a lasting impact on future research. In addition to the meteorological observations made using WFIP2 equipment, the data set includes archives from external data sources including our many wind industry partners. The collocation of wind turbine observations with surrounding and in situ meteorological observations make the data set uniquely valuable. We know of no other dataset covering such a large region of wind energy production for such a prolonged period of time.

Analysis of our observations formed the basis for the advancement of our understanding of complex near-surface flows that affect turbine hub-height wind speeds, and subsequent WFIP2 model development efforts. The analysis and development phase of WFIP2 was guided by our observational event log, as every recorded event included an analysis of its importance to wind power forecasting.

Validation of model development generally proceeded from examination of specific case studies to larger blocks of time and eventually to the creation of full-year retrospective forecast datasets. As with the observations, all model output from the project was made publicly available via DOE's DAP. The Vaisala's teams validation of the WFIP2 retrospective forecasts showed modest improvements in forecast skill, primarily during wintertime.

To facilitate industry uptake of increasingly probabilistic wind power forecasts, particularly those presenting event-based probabilities, a prototype decision support tool was developed and shared with industry partners.

A major strength of WFIP2 was the level of cooperation between DOE and NOAA laboratories and industry. Even though much of the research in the project was fundamental in nature, it was continuously guided by the needs of forecasting, and as a result the time between research and implementation in operations was greatly reduced, with several model developments already running in NOAA's operational model. In addition, decision support tools were developed and presented to the wind energy industry. These tools combine forecasts and forecast uncertainty in an interface designed to facilitate action at probabilistic levels appropriate to specific risk/reward levels.

The WFIP2 team (and the Vaisala Team within it) met many of its original goals while responding well to the challenges of a prolonged field campaign and the uncertainties of algorithm development. The organization of the project by functional team rather than by institution, and regular steering committee meetings, facilitated the success of the project. Subordination of organization affiliations to team goals and effort was evident throughout.

In addition to efforts from within the team, WFIP2 benefited from external programs like the Data Archive and Portal and the Argonne Leadership Computing Facility, without which we would not have been able to process and provide publicly the volume of data that we required to make significant model advancements.

Complex terrain both concentrates wind energy into highly developable areas of excellent wind resource and also drives numerous meteorological phenomena that make wind energy forecasting especially

challenging. However, the economics of the wind energy development are leading the industry to build more projects in complex terrain. While this study represents an advancement in our understanding and modeling of the winds that drive such projects, we hope that it is seen as a step in an ongoing process to improve the state of forecasting in these areas.

References

Banakh, V. A., and I. N. Smalikho, 1997: Estimation of the turbulent energy dissipation rate from pulsed Doppler lidar data. *Atmos. Oceanic Opt.*, **10,** 957–965.

Beare, R. J., 2013: A length scale defining partially-resolved boundary-layer turbulence simulations. *Bound.-Layer Meteor.*, **151**, 39–55, doi: https://doi.org/10.1007/s10546-013-9881-3.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi: https://doi.org/10.1175/MWR-D-15-0242.1.

Berg, L. K., M. Pekour, and D. Nelson, 2012: Description of the Columbia Basin Wind Energy Study (CBWES). Pacific Northwest National Laboratory Tech. Rep. PNNL-22036, 24 pp, http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-22036.pdf.

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

Bodini, N., J. K. Lundquist, and R. K. Newsom, 2018: Estimation of turbulence dissipation rate and its variability from sonic anemometer and wind Doppler lidar during the XPIA field campaign, *Atmos. Meas. Tech.*, **11,** 4291–4308, doi: https://doi.org/10.5194/amt-11-4291-2018.

Bougeault, P., and P. Lacarrere, 1989: Parameterization of orography-induced turbulence in a mesobeta-scale model. *Mon. Wea. Rev.*, **117**, 1872–1890, https://doi.org/10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2.

Boutle, I. A., J. E. J. Eyre, and A. P. Lock, 2014: Seamless stratocumulus simulation across the turbulent gray zone. *Mon. Wea. Rev.*, **142**, 1655–1668, doi: https://doi.org/10.1175/MWR-D-13-00229.1.

Bowen, A., and N. Mortensen, 1996: Exploring the limits of WAsP: The wind atlas analysis and application program. *Proc. European Union Wind Energy Conf.,* Goteborg, Sweden, European Wind Energy Association, 584–587.

Bradley, E. F. 1980. An experimental study of the profiles of wind speed, shearing stress and turbulence at the crest of a large hill. *Quart. J. Roy. Meteor. Soc.*, **106,** 101–124, doi: https://doi.org/10.1002/qj.49710644708.

Chan, P. W., and A. M. Shao, 2007: Depiction of complex airflow near Hong Kong International Airport using a Doppler lidar with a two-dimensional wind retrieval technique. *Meteor. Z.*, **16**, 491–504, doi: http://doi.org/10.1127/0941-2948/2007/0220.

Ching, J., R. Rotunno, M. LeMone, A. Martilli, B. Kosovic, P. A. Jimenez, and J. Dudhia, 2014: Convectively induced secondary circulations in fine-grid mesoscale numerical weather prediction models. *Mon. Wea. Rev.*, **142**, 3284–3302, doi: https://doi.org/10.1175/MWR-D-13-00318.1.

Clements, W. E., J. A. Archuleta, and P. H. Gudiksen, 1989: Experimental design of the 1984 ASCOT field study. *J. Appl. Meteor.*, **28,** 405–413.

Coulter, R. L., and T. J. Martin, 1996: Effects of stability on the profiles of vertical velocity and its variance in katabatic flow. *Bound.-Layer Meteor.*, **81,** 23–33.

Cuxart, J. C., and Coauthors, 2000: Stable Atmospheric Boundary-Layer Experiment in Spain (SABLES-98): A report. Bound.-Layer Meteor., **96**, 337–370, doi: *https://doi.org/10.1023/A:1002609509707*.

de Franceschi, M., D. Zardi, M. Tagliazucca, and F. Tampieri, 2009: Analysis of second-order moments in the surface layer turbulence in an Alpine valley. *Quart. J. Roy. Meteor. Soc.*, **135,** 1750–1765, doi: https://doi.org/10.1002/qj.506.

Doran, J. C., J. D. Fast, and J. Horel, 2002: The VTMX 2000 Campaign. *Bull. Amer. Meteor. Soc.*, **83**, 537–551.

Doviak, R. J., and D. S. Zrnić, 1993: *Doppler Radar and Weather Observations.* 2nd ed. Academic Press, 562 pp.

Efstathiou, G. A., and R. J. Beare, 2015: Quantifying and improving sub-grid diffusion in the boundary-layer grey zone. *Quart. J. Roy. Meteor. Soc.*, **141**, 3006–3017, doi:10.1002/qj.2585.

Efstathiou, G. A., R. J. Beare, S. Osborne, and A. P. Lock, 2016: Grey zone simulations of the morning convective boundary layer development. *J. Geophys. Res. Atmospheres*, **121**, 4769–4782, doi:10.1002/2016JD024860.

Fernando, H. J. S., 2013: *Handbook of Environmental Fluid Dynamics.* CRC Press, Boca Raton, 624 pp, doi: https://doi.org/10.1201/b14241.

Fernando, H. J. S., and E. R. Pardyjak, 2013: Field studies delve into the intricacies of mountain weather. *Eos, Trans. Amer. Geophys. Union*, **94**, 313–315, doi: https://doi.org/10.1002/2013EO360001.

Fernando, H. J. S., and Coauthors, 2015: The MATERHORN—Unraveling the intricacies of mountain weather. *Bull. Amer. Meteor. Soc.*, doi: https://doi.org/10.1175/BAMS-D-13-00131.1.

Frehlich, R., S. M. Hannon, and S. W. Henderson, 1998: Coherent Doppler lidar measurements of wind field statistics. Bound.-Layer Meteor., 86, 233–256, doi: https://doi.org/10.1023/A:1000676021745.

Frehlich, R., Y. Meillier, M. L. Jensen, and B. Balsley, 2006: Measurements of boundary layer profiles in an urban environment. J. Appl. Meteor. Climatol., 45, 821–837, doi: https://doi.org/10.1175/JAM2368.1.

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 farms as parameterized in a mesoscale NWP model. Mon. Wea. Rev., 140, 3017–3038, doi: https://doi.org/10.1175/MWR-D-11-00352.1.

Fitch, A. C., J. K. Lundquist, and J. B. Olson, 2013: Mesoscale influences of wind farms throughout a diurnal cycle. Mon. Wea. Rev., 141, 2173–2198, doi: https://doi.org/10.1175/MWR-D-12-00185.1.

Garratt, J. R., 1992: *The Atmospheric Boundary Layer.* Cambridge University Press, 316 pp.

Grimit, E., 2019: The second Wind Forecast Improvement Project (WFIP2) decision support tools. *Bull. Amer. Meteor. Soc.*, in preparation.

Grubišić, V., J.D. Doyle, J. Kuettner, S. Mobbs, R.B. Smith, C.D. Whiteman, R. Dirks, S. Czyzyk, S.A. Cohn, S. Vosper, M. Weissmann, S. Haimov, S.F. De Wekker, L.L. Pan, and F.K. Chow, 2008: The terrain-induced rotor experiment. *Bull. Amer. Meteor. Soc.,* **89**, 1513–1534, https://doi.org/10.1175/2008BAMS2487.1.

Honnert, R., and V. Masson, 2014: What is the smallest physically acceptable scale for 1D turbulence schemes? *Front. Earth Sci.*, **2**, https://doi.org/10.3389/feart.2014.00027.

Hunt, J. C. R., Y. Feng, P. F. Linden, M. D. Greenslade, and S. D. Mobbs, 1997: Low Froude number stable flows past mountains. *Il Nuovo Cimento*, **20C,** 261–271.

Ito, J., H. Niino, M. Nakanishi, and C.-H. Moeng, 2015: An Extension of the Mellor–Yamada Model to the Terra Incognita Zone for Dry Convective Mixed Layers in the Free Convection Regime. *Bound.-Layer Meteor.*, **157**, 23–43, doi:10.1007/s10546-015- 0045-5.

Jiménez, P. A., J. Dudhia, J. F. González-Rouco, J. Navarro, J. P. Montávez, and E. García-Bustamante, 2012: A revised scheme for the WRF surface layer formulation. *Mon. Wea. Rev.*, **140**, 898–918, https://doi.org/10.1175/MWR-D-11-00056.1.

Kaimal, J. C., and J. J. Finnigan, 1994: *Atmospheric Boundary Layer Flows: Their Structure and Measurement.* Oxford University Press, 289 pp.

Kaimal, J. C., J. C. Wyngaard, Y. Izumi, and O. R. Coté, 1972: Spectral characteristics of surface-layer turbulence. *Quart. J. Roy. Meteor. Soc.*, **98,** 563–589, doi: https://doi.org/10.1002/qj.49709841707.

Kaimal, J. C., J. C. Wyngaard, D. A. Haugen, O. R. Coté, Y. Izumi, S. J. Caughey, and C. J. Readings, 1976: Turbulence structure in the convective boundary layer. *J. Atmos. Sci.*, **33**, 2152–2169, doi: https://doi.org/10.1175/1520-0469(1976)033<2152:TSITCB>2.0.CO;2.

Krishnamurthy, R. , Calhoun, R. , Billings, B. and J. Doyle, 2011: Wind turbulence estimates in a valley by coherent Doppler lidar. *Meteor. Appl.*, **18,** 361-371, doi: http://doi.org/10.1002/met.263.

Krishnamurthy, R., A. Choukulkar, R. Calhoun, J. Fine, A. Oliver, and K. S. Barr, 2013: Coherent Doppler lidar for wind farm characterization. *Wind Energy*, **16**, 189–206, doi: https://doi.org/10.1002/we.539.

Lee, J. C. Y. and J. K. Lundquist, 2017: Evaluation of the wind farm parameterization in the Weather Research and Forecasting model (version 3.8.1) with meteorological and turbine power data, *Geosci. Model Dev.*, **10**, 4229-4244, doi: https://doi.org/10.5194/gmd-10-4229-2017.

Marquis, M., J. Wilczak, M. Ahlstrom, J. Sharp, A. Stern, J. Smith, and S. Calvert, 2011: Forecasting the wind to reach significant penetration levels of wind energy. *Bull. Amer. Meteor. Soc.*, **92**, 1159–1171, https://doi.org/10.1175/2011BAMS3033.1.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, doi: https://doi.org/10.1256/003590002320603584.

McCaffrey, K., L. Bianco, and J. M. Wilczak, 2017: Improved observations of turbulence dissipation rates from wind profiling radars. *Atmos. Meas. Tech.*, **10**, 2595–2611, doi: https://doi.org/10.5194/amt-10-2595-2017.

McCaffrey, K., and Coauthors, 2019: Identification and characterization of persistent cold pool events from temperature and wind profilers in the Columbia river basin. J*. Appl. Meteor. Climatol.*, in review.

McMillen, R. T., 1988: An eddy correlation technique with extended applicability to non-simple terrain. *Bound.-Layer Meteor.,***43,** 231–245, doi: https://doi.org/10.1007/BF00128405.

Mellor, G. L., and T. Yamada, 1974: A hierarchy of turbulence closure models for planetary boundary layers. *J. Atmos. Sci.*, **31**, 1791–1806, doi: https://doi.org/10.1175/1520-0469(1974)031<1791:AHOTCM>2.0.CO;2.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875, doi:10.1029/RG020i004p00851.

Monin, A. S., and A. M. Yaglom 1975: Statistical fluid mechanics: Mechanics of turbulence – Vol 2, The MIT Press, Cambridge, 896 pp.

Muñoz-Esparza, D., B. Kosović, J. Mirocha, and J. van Beeck, 2014: Bridging the transition from mesoscale to microscale turbulence in numerical weather prediction models. *Bound.-Layer Meteor.*, **153**, 409–440, doi: https://doi.org/10.1007/s10546-014-9956-9.

Muñoz-Esparza, D., B. Kosović, J. van Beeck, and J. Mirocha, 2015: A stochastic perturbation method to generate inflow turbulence in large-eddy simulation models: Application to neutrally stratified atmospheric boundary layers. *Phys. Fluids*, **27**, 035102, doi: https://doi.org/10.1063/1.4913572.

Nakanishi, M., 2001: Improvement of the Mellor-Yamada turbulence closure model based on large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378, doi:10.1023/A:1018915827400.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, doi: https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Jpn. Ser II*, **87**, 895–912, doi: https://doi.org/10.2151/jmsj.87.895.

Olson, J. B. and Coauthors, 2019: Improving wind energy forecasting through numerical weather prediction model development, *Bull. Amer. Meteor. Soc.*, in review.

Orgill, M. M., and R. I. Schreck, 1985: An overview of the ASCOT multi-laboratory field experiments in relation to drainage winds and ambient flow. *Bull. Amer. Meteor. Soc.*, **66**, 1263–1277.

Pahlow, M., M. B. Parlange, and F. Porte-Agel, 2001: On Monin-Obukhov similarity in the stable atmospheric boundary layer. *Bound.-Layer Meteor*, **99,** 225–248, doi: https://doi.org/10.1023/A:1018909000098.

Panofsky, H. A., and J. A. Dutton, 1984: *Atmospheric Turbulence.* John Wiley and Sons, New York, 397 pp.

Papanastasiou, D. K., D. Melas, and I. Lissaridis, 2010: Study of wind field under sea breeze conditions; an application of WRF Model. *Atmos. Res.*, **98**, 102–117, doi: https://doi.org/10.1016/j.atmosres.2010.06.005.

Petersen, E. L., N. G. Mortensen, L. Landberg, J. Højstrup, and H. P. Frank, 1998: Wind power meteorology. Part I: Climate and turbulence. *Wind Energy*, **1**, 2–22, doi: https://doi.org/10.1002/(SICI)1099-1824(199809)1:1<2::AID-WE15>3.0.CO;2-Y.

Piper, M., and J. K. Lundquist, 2004: Surface layer turbulence measurements during a frontal passage. *J. Atmos. Sci.*, **61,** 1768–1780, doi: https://doi.org/10.1175/1520-0469(2004)061<1768:SLTMDA>2.0.CO;2.

Poulos, G. S., and Coauthors, 2002: CASES-99: A comprehensive investigation of the stable nocturnal boundary layer. *Bull. Amer. Meteor. Soc*, **83,** 555–581, doi: https://doi.org/10.1175/1520-0477(2002)083<0555:CACIOT>2.3.CO;2.

Redfern, S., J. B. Olson, J. K. Lundquist, and C. T. Clack, 2019: Incorporation of the Rotor-Equivalent Wind Speed into the Weather Research and Forecasting Model's Wind Farm Parameterization. *Mon. Wea. Rev.,* **in press**, https://doi.org/10.1175/MWR-D-18-0194.1.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649-667, doi: https://doi.org/10.1002/qj.49712656313.

Rotach, M. W., and D. Zardi, 2007: On the boundary-layer structure over highly complex terrain: Key findings from MAP. *Quart. J. Roy. Meteor. Soc.*, **133**, 937–948, doi: https://doi.org/10.1002/qj.71.

Roulston, M. S., D. T. Kaplan, J. Hardenberg, and L. A. Smith, 2003: Using medium-range weather forecasts to improve the value of wind energy production. *Renewable Energy*, **28**, 585–602, doi: https://doi.org/10.1016/S0960-1481(02)00054-X.

Schreck, S, J. Lundquist, and W. Shaw, 2008: U.S. Department of Energy Workshop Report—Research needs for wind resource characterization. NREL Rep. TP-500-43521, 116 pp.

Sharp, J., and C. F. Mass, 2002: Columbia Gorge gap flow: Insights from observational analysis and ultra-high resolution simulation. *Bull. Amer. Meteor. Soc.*, **83**, 1757–1762.

Sharp, J., and C. F. Mass, 2004: Columbia Gorge gap winds: Their climatological influence and synoptic evolution. *Wea. Forecasting*, **19**, 970–992, doi: https://doi.org/10.1175/826.1.

Shaw, W. J. and Coauthors, 2019: The second Wind Forecast Improvement Project (WFIP2): General Overview. *Bull. Amer. Meteor. Soc.* Accepted.

Sheppard, P. A. 1956. Airflow over mountains. *Quart. J. Roy. Meteor. Soc.*, **82,** 528–529.

Shin, H. H., and J. Dudhia, 2016: Evaluation of PBL parameterizations in WRF at subkilometer grid spacings: Turbulence statistics in the dry convective boundary layer. *Mon. Wea. Rev.*, **144**, 1161–1177, doi: https://doi.org/10.1175/MWR-D-15-0208.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Weather Rev.*, **132**, 3019–3032.

Skamarock, W. C., and J. B. Klemp, 2008: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J. Comput. Phys.*, **227**, 3465–3485, doi:10.1016/j.jcp.2007.01.037.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research WRF version 2. NCAR Tech Note NCAR/TN-468+STR, 88 pp.

Smagorinsky, J., 1990: Some historical remarks on the use of non-linear viscosities in geophysical models. *Proc. Int. Workshop Large Eddy Simul. We Stand Eds*, 1-3.

Smith, N. H., 2018: A sensitivity study on wind ramp events in the Columbia river basin. Ph.D. dissertation, Texas Tech University, 181 pp, https://hdl.handle.net/2346/82690.

Stiperski, I., and M. W. Rotach, 2016: On the measurement of turbulence over complex mountainous terrain. *Bound.-Layer Meteor.*, **159**, 97–121, doi: https://doi.org/10.1007/s10546-015-0103-z.

Večenaj, Ž., D. Belušić, and B. Grisogono, 2010: Characteristics of the near-surface turbulence during a Bora event. *Ann. Geophys*, **28,** 155–163, doi: https://doi.org/10.5194/angeo-28-155-2010.

Wagner, R., B. Cañadillas, A. Clifton, S. Feeney, N. Nygaard, M. Poodt, C. St Martin, E. Tüxen, and J. W. Wagenaar, 2014: Rotor equivalent wind speed for power curve measurement–comparative exercise for IEA Wind Annex 32. *Journal of Physics: Conference Series*, **524**, 012108.

Warner, T. T., 2010: *Numerical weather and climate prediction*. Cambridge University Press, Cambridge, 522 pp, https://doi.org/10.1017/CBO9780511763243.

Whiteman, C. D., and Coauthors, 2008: Metcrax 2006. *Bull. Amer. Meteor. Soc.*, **89**, 1665–1680, https://doi.org/10.1175/2008BAMS2574.1.

Wilczak, J. M., and Coauthors, 2015: The Wind Forecast Improvement Project (WFIP): A public–private partnership addressing wind energy forecast needs. *Bull. Amer. Meteor. Soc.,* **96**, 1699–1718, https://doi.org/10.1175/BAMS-D-14-00107.1.

Wilczak, J. M., and Coauthors, 2019: The second Wind Forecast Improvement Project (WFIP2): Observational field campaign. *Bull. Amer. Meteor. Soc.*, accepted.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.*, **8**, 209–219, doi: https://doi.org/10.1017/S1350482701002092.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Wood, N., 1995: The onset of separation in neutral, turbulent flow over hills. Bound.-Layer Meteor., 76, 137–164, doi: https://doi.org/10.1007/BF00710894.

Wood, N., 2000: Wind flow over complex terrain: A historical perspective and the prospect for large-eddy modelling. *Bound.-Layer Meteor.*, **96**, 11–32, doi: https://doi.org/10.1023/A:1002017732694.

Wyngaard, J.C., 2004: Toward numerical modeling in the “terra incognita”. *J. Atmos. Sci.*, **61,** 1816–1826, https://doi.org/10.1175/1520-0469(2004)061<1816:TNMITT>2.0.CO;2.

Yamada, T., 1975: The critical Richardson number and the ratio of the eddy transport coefficients obtained from a turbulence closure model. *J. Atmos. Sci*, **32,** 926–933, doi: https://doi.org/10.1175/1520-0469(1975)032<0926:TCRNAT>2.0.CO;2.

Yang, B., Y. Qian, L. K. Berg, P.-L. Ma, S. Wharton, V. Bulaevskaya V., 2017: Sensitivity of turbine-height wind speeds to parameters in planetary boundary-layer and surface-layer schemes in the Weather Research and Forecasting model. *Bound.-Layer Meteor*. **162**(1):117–142, https://doi.org/10.1007/s10546-016-0185-2.

Zeman, O., and N. O. Jenssen, 1987: Modification to turbulence characteristics in flow over hills. *Quart. J. Roy. Meteor. Soc.*, **113,** 55–80, doi: https://doi.org/10.1002/qj.49711347505.

Zhang, K., and Coauthors, 2001: Numerical investigation of boundary layer evolution and nocturnal low-level jets: Local versus non-local PBL schemes. *Environ. Fluid Mech.,* **1**, 171.

List of Acronyms

A2e: Atmosphere to Electrons

ABL: Atmospheric Boundary Layer

AGL: Above Ground Level

ARW: Advanced research WRF

ASCOT: Atmospheric Studies in Complex Terrain

ASL: Above Sea Level

ASOS: Automated Surface Observing System

AUC: Area Under the Relative Operating Characteristic Curve

AWOS: Automated Weather Observing System

BAMS: Bulletin of the American Meteorological Society

BPA: Bonneville Power Administration

CU: University of Colorado

DAP: Data Archive and Portal

DOE: Department of Energy

DST: Decision Support Tool

ECMWF: European Centre for Medium-range Weather Forecasting

EDR: Eddy Dissipation Rate

EERE: Energy Efficiency and Renewable Energy

EnKF: Ensemble Kalman Filter

ESRL: Earth Systems Research Laboratory

ETS: Equitable Threat Score

FAR: False Alarm Ratio

FHT: Flat and Homogeneous Terrain

HQ: Headquarters

HRRR: High Resolution Rapid Refresh model

NCEP: National Center for Environmental Prediction

NOAA: National Oceanic and Atmospheric Administration

HRRRNEST: HRRR WFIP2 Nest

IC: Initial Conditions

JD: Julian Day

JJA: June-July-August

LBC: Lateral Boundary Conditions

LES: Large Eddy Scale

LIDAR: Light Detection and Ranging

LMCO: Lockheed Martin Inc.

LT: Local Time

MAE: Mean Absolute Error

MAP: Mesoscale Alpine Programme

MATERHORN: Mountain Terrain Atmospheric Modeling and Observations Program

MOS: Model Output Statistics

MOST: Monin-Obhukov Similarity Theory

MSL: Mean Sea Level

MYNN: Mellor-Yamada-Nakanishi-Niino

NCAR: National Center for Atmospheric Research

NCEP: National Centers for Environmental Prediction

NED: National Elevation Dataset

NOAA: National Oceanic and Atmospheric Administration

NWP: Numerical Weather Prediction

NWS: National Weather Service

PBL: Planetary Boundary Layer

PDT: Pacific Daylight Time

PGE: Portland General Electric

PIRT: Phenomena Identification and Ranking Table

PNNL: Pacific Northwest National Laboratories

POD: Probability of Detection

PS: Physics Site

RANS: Reynolds Averaged Navier-Stokes

REWS: Rotor Equivalent Wind Speed

T-REX: Terrain Induced Rotor Experiment

RH: Relative Humidity

RIX: Ruggedness Index

RMS: Root Mean Square

RMSE: Root Mean Square Error

ROC: Relative Operating Characteristic

RR: Ridge Regression

SABLES: Stable Atmospheric Boundary-Layer Experiment

SNR: Signal-Noise Ratio

SODAR: Sound Detection and Ranging

TKE: Turbulent Kinetic Energy

TTU: Texas Tech University

UND: University of Notre Dame

UQ: Uncertainty Quantification

UTC: Coordinated Universal Time

VTMX: Vertical Transport and Mixing Experiment

WFIP1: Wind Forecast Improvement Project 1

WFIP2: Wind Forecast Improvement Project 2

WFP: Wind Farm Parameterization

WRF: Weather Research and Forecasting model