Measurements and Model Improvement: Insight into NWP Model Error Using Doppler Lidar and Other WFIP2 Measurement Systems

ROBERT M. BANTA,a,b YELENA L. PICHUGIN,a,b W. ALAN BREWER,b KELLY A. BALMES,a,c BIANCA ADLER,a,d JOSEPH SEDLAR,a,c LISA S. DARBY,a DAVID D. TURNER,h JAYMES S. KENYON,a,c EDWARD J. STROBACH,a,h BRIAN J. CARROLL,a,b JUSTIN SHARP,h MARK T. STOELINGA,j JOEL CLINE,i AND HARINDRA J. S. FERNANDOg

a CIRES, University of Colorado Boulder, Boulder, Colorado
b NOAA/Chemical Sciences Laboratory, Boulder, Colorado
c NOAA/Tropical Program Coordinator, Silver Spring, Maryland
d NOAA/Physical Systems Laboratory, Boulder, Colorado
e NOAA/Global Sciences Laboratory, Boulder, Colorado
f LDWX LLC, Boulder, Colorado
g Vaisala, Seattle, Washington
h NOAA/Global Monitoring Laboratory, Boulder, Colorado
i NOAA/Global Monitoring Laboratory, Boulder, Colorado
j CIRES, University of Colorado Boulder, Boulder, Colorado

(Manuscript received 26 March 2023, in final form 2 September 2023, accepted 7 September 2023)

ABSTRACT: Doppler-lidar wind-profile measurements at three sites were used to evaluate NWP model errors from two versions of NOAA’s 3-km-grid HRRR model, to see whether updates in the latest version 4 reduced errors when compared against the original version 1. Nested (750-m grid) versions of each were also tested to see how grid spacing affected forecast skill. The measurements were part of the field phase of the Second Wind Forecasting Improvement Project (WFIP2), an 18-month deployment into central Oregon–Washington, a major wind-energy-producing region. This study focuses on errors in simulating marine intrusions, a summertime, 600–800-m-deep, regional sea-breeze flow found to generate large errors. HRRR errors proved to be complex and site dependent. The most prominent error resulted from a premature drop in modeled marine-intrusion wind speeds after local midnight, when lidar-measured winds of greater than 8 m s\(^{-1}\) persisted through the next morning. Large negative errors were offset at low levels by positive errors due to excessive mixing, complicating the interpretation of model “improvement,” such that the updates to the full-scale versions produced mixed results, sometimes enhancing but sometimes degrading model skill. Nesting consistently improved model performance, with version 1’s nest producing the smallest errors overall. HRRR’s ability to represent the stages of sea-breeze forcing was evaluated using radiation budget, surface-energy balance, and near-surface temperature measurements available during WFIP2. The significant site-to-site differences in model error and the complex nature of these errors mean that field-measurement campaigns having dense arrays of profiling sensors are necessary to properly diagnose and characterize model errors, as part of a systematic approach to NWP model improvement.

SIGNIFICANCE STATEMENT: Dramatic increases in NWP model skill will be required over the coming decades. This paper describes the role of major deployments of accurate profiling sensors in achieving that goal and presents an example from the Second Wind Forecast Improvement Program (WFIP2). Wind-profile data from scanning Doppler lidars were used to evaluate two versions of HRRR, the original and an updated version, and nested versions of each. This study focuses on the ability of updated HRRR versions to improve upon predicting a regional sea-breeze flow, which was found to generate large errors by the original HRRR. Updates to the full-scale HRRR versions produced mixed results, but the finer-mesh versions consistently reduced model errors.

KEYWORDS: Lidars/Lidar observations; Model evaluation/performance; Experimental design; Sea breezes; Regional effects; Renewable energy

1. Introduction

Improving NWP forecast-model skill is a major atmospheric-science research aspiration. Measurements play a key role in successfully achieving that goal, needed at least to evaluate whether a new version of a model has reduced errors versus the old version. Advances in measurement capabilities and approaches are needed for more rapid model-skill improvement.

The Second Wind Forecast Improvement Project (WFIP2) was a measurement-modeling research program designed to exploit recent progress in Doppler lidar and other measurement capabilities. The field-measurement phase took place in
the complex terrain of the Columbia River Valley (CRV) of central Oregon–Washington (Fig. 1a) during an 18-month campaign. Designed to enhance the quality of atmospheric information available to the wind-energy (WE) community, the instrument deployment also had a model verification and improvement objective (Shaw et al. 2019), in addition to its goal of better understanding weather systems that control the wind profile in the lowest 200 m of atmosphere (Banta 2022). The primary model evaluated was NOAA/NCEP’s High-Resolution Rapid Refresh (HRRR) operational NWP forecast model. Although HRRR is continental in scale, many improvements would address general modeling issues applicable to other model scales, including global. WFIP2 verification studies to date have assessed HRRR model errors averaged over periods of weeks to a year or longer (Pichugina et al. 2019, 2020, 2022; Olson et al. 2019a; Bianco et al. 2019). These studies found significant horizontal variability in the nature of the model errors from site to site, even within the same basin. A significant implication is, “if only one of the sites had been instrumented for model evaluation, ... a very different picture of the nature of the model errors” would result, depending on which site had been instrumented (Pichugina et al. 2019).

Another goal of WFIP2 was to investigate large model errors and the specific flow phenomena or regimes that produce them. One such large-error phenomenon is the diurnal marine intrusion, a regional, warm-season sea-breeze wind surge that begins at the coast in the morning and then penetrates into the WFIP2 study region by late afternoon (see Fig. 2 and animation in the online supplemental material). The present study is the third in a series on the marine intrusion. The first (Banta et al. 2020) used wind-profile analyses from three scanning Doppler lidars (Fig. 1) to define this flow type, finding striking similarities among the eight cases identified during summer 2016. This study also used the lidar-profile measurements to evaluate errors in the operational HRRR at the time (HRRR-NCEP version 1), revealing an unexpected (perhaps even remarkable) similarity in model-error behavior—an error signature—from case to case among the eight occurrences (Banta et al. 2018b, 2020). The second study (Banta et al. 2021) looked at the major daily wind-flow patterns or regimes encountered during June–July–August 2016, which included those driven primarily by the large-scale pressure gradient across the Cascade Mountain Range and those driven mostly by differential heating effects, such as land–sea

FIG. 1. (a) Study area in the Columbia River Valley: locations of scanning Doppler lidars at Wasco, Arlington, and Boardman, OR, denoted by yellow circles, and NWS sites shown in white capital letters: Astoria (AST), Portland airport (PDX), Troutdale (TTD), the Dalles (DLS), and Hermiston (HRI). (b) Enlargement of study area, top portion showing Google Earth map of the WFIP2 study area, with the locations of Doppler lidars. Surrounding wind farms indicated by clusters of orange dots (each dot represents a turbine) and white line indicates a WSW–ENE transect of the study region. The bottom portion of (b) shows terrain elevation along this line, where lidar locations are indicated by black stars. (c) Domains of NOAA’s RAP and HRRR operational models (white and orange boxes) and the smaller domains of the provisional HRRR and HRRR-nest runs used in the present study [green boxes; from Olson et al. (2019a)]. (d) Study area terrain as represented in 3-km HRRR, and (e) terrain as in 750-m HRRR-nest. For enlargements of (c) and (d), see the supplemental material.
contrast (Staley 1959) forcing. The marine-intrusion regime in the latter category showed the smallest case-to-case variation and the largest model errors of all the regimes, a significant finding being that the larger the role of differential heating in generating the flow, the larger the model errors.

In this, the third study, we address model-improvement issues for WFIP2 marine-intrusion cases. We examine reasons for the large marine-intrusion errors and use available measurements to evaluate the effectiveness of three rounds of HRRR updates implemented since summer 2016 in reducing the large errors. We also evaluate the effects of decreasing the HRRR grid interval from 3 km to 750 m. We compare lidar wind profiles against simulated profiles by the first version of HRRR (HRRR-v1, which we will refer to here as simply “v1”) and the latest version.

Fig. 2. (a) Wind-generated power (GW, from BPA website) for 6 days in August 2016, showing the dramatic 24-h periodicity that drew our attention to these flows, occurring on 4 consecutive days (14–17 Aug). (b) Doppler-lidar time–height cross sections of wind speed on 17 Aug with a several-hundred-meter-deep burst of \( >10 \) m s\(^{-1}\) winds traveling from west to east in late afternoon, and remaining strong until the next afternoon (Banta et al. 2020). Vertical axis is height AGL (km), horizontal axis is hour (UTC), and the color bar is wind speed (m s\(^{-1}\)). (c) Time series of wind speed at three heights (50, 100, and 150 m AGL), corresponding to rotor-layer wind speeds in the cross sections in (b). Nighttime hours indicated by dark bar along abscissa of time series panels.
flow phenomena as recommended by Banta et al. (2020) and to adequately capture the mix of weather types occurring over the annual cycle. WFIP2 lasted 18 months, but three scanning Doppler lidars were not enough to adequately sample the spatial variability as required. Even studying a relatively straightforward regional sea breeze over the limited CRV region, a density of perhaps 3 times or more of such sensors would be required for adequate coverage over this region.

b. Quantifying skill improvement: Role of measurement accuracy

Errors are difference quantities between modeled and measured variables, two quantities that are presumably close in value, so accuracy of measurement is essential to calculate useful model errors. Finding error differences—differences of differences—among measurement sites (Pichugina et al. 2019, 2020), among weather regimes (Banta et al. 2021), or, as here, among model versions, requires even greater measurement accuracy. How accurate is a subject for further study and debate, but for winds, the tolerance is most likely on the order of 0.1 m s⁻¹ (Banta et al. 2013a).

This value is supported by a recent WFIP2 study (Pichugina et al. 2022), in which collocated wind sensors at the three measurement sites (Fig. 1) evaluated annual error statistics for 15-min, 80-m wind speeds predicted by the two HRRR versions (and their nests) considered here: v1, the original version, and v4, which incorporated many model physics and other updates. Different sensors, including Doppler sodars and different types of Doppler lidar, produced different values of the measured basic mean wind speed at each site, as expected, typically varying by 0.1–0.2 m s⁻¹, but at times exceeding 0.4 m s⁻¹. Model errors were centered around 3 m s⁻¹ (RMSE), in agreement with previous WFIP2 studies (see section 3c). For each model version tested and for each site, the sensor-to-sensor variability among calculated errors amounted to ~10% of this value, or 0.3 m s⁻¹. Thus, when comparing the changes in model errors (Δerror) from v1 to v4, each instrument at a given site found different Δerror values. The instrument-to-instrument differences in ΔRMSE, for example, were ~0.1 m s⁻¹—but some were as large as 0.3 m s⁻¹. Mostly this meant that one instrument indicated a larger increase or decrease in skill than the other, but in some cases, the model v1-to-v4 error difference grew with respect to one sensor’s measurements and declined with respect to another—one instrument would show that the updates improved model skill, and the other would show that they degraded skill. Because model updates that increase model error should be discarded, and those that decrease error retained, it becomes important to know which sensor’s conclusion to believe: hence the importance to model improvement of understanding measurement accuracy.

An example of a measurement issue is measurement bias. As just shown, a small measurement bias of even 0.1 m s⁻¹ or so can give a false indication; it can prejudice the comparison in favor of one model version, when an unbiased-measurement comparison might favor the other. Such measurement biases could result, for example, from hardware distortions or misalignments, violation of sampling assumptions, weak signal, tilting of the vertical beam
off zenith for Doppler-beam swinging (DBS) profiling devices, complex-terrain effects, scattering targets that do not drift with the wind (e.g., Geerts and Miao 2005), and others. Bias and other measurement uncertainties are an important and active area of investigation for remote sensing systems, especially for systems that may be unattended for several weeks or months at a time, and those findings will be essential background information for serious model-improvement efforts.

**c. Role of marine-intrusion studies**

NWP forecast models must at least get the dry processes right, including many of those deemed for decades to be necessary for NWP model improvement in general, such as subresolution-scale (SRS) transfer and mixing, radiation, energy exchanges at Earth’s surface, and other related processes (e.g., Seaman 2000; Dabberdt et al. 2004; Jakob 2010). Finding conditions where the dry processes can be isolated for model evaluation may be difficult because in NWP models all processes are intricately linked, so such isolation seems unlikely.

Under certain summertime conditions during WFIP2, however, one type of flow, the marine intrusion studied here, evolves in the same manner from case to case (Banta et al. 2020, 2021), despite significant variation in the 700-hPa winds, an indicator of large-scale variability, among the occurrences. Such constancy of flow development indicates separation of the essential marine-intrusion dynamics from external, larger-scale influences. Not only did the flow evolve in the same way, but the HRRR model errors also showed the characteristic diurnal “signature” from case to case, as mentioned, in the form of premature drops of the strong intrusion flow when a marine intrusion occurred, HRRR made the same mistakes each time. These days were dry and cloudless, except for some thin cirrus on individual days. We concluded from this behavior that the dominant physical processes generating this flow were dry, that they all occurred within a limited region or purlieu of a few 100 km about the study region where the errors were measured (Banta et al. 2020), and that the same was true in the model—the diurnal flow evolution and accompanying errors were all generated within this purlieu, independently of larger-scale modeled processes and their associated imported errors. As this flow is a regional sea breeze, the basic physics are well known.

The errors were discovered using v1, the version of HRRR being used operationally during WFIP2 (Pichugina et al. 2019; Banta et al. 2020). The current operational version is v4, which has undergone three rounds of updates (Dowell et al. 2022). Internal verification studies have indicated that these updates improved model skill in general. Also important, though, is how these updates affect model skill in simulating large-error atmospheric phenomena. Here, we investigate whether the v4 updates have reduced the large errors generated by v1 specifically for marine intrusions.

These WFIP2 studies focus on smaller-scale features and processes. Studies using coarser-resolution, global-scale models, however, in which near-surface processes are heavily parameterized, have revealed sensitivity to the small-scale exchanges within the atmospheric boundary layer (ABL), as represented in these models. For example, “reducing the diffusion in stable layers situated near the surface has a direct effect on the amplitude of the planetary-scale standing waves,” and on other large-scale performance metrics, such as “the root mean square of the geopotential height at 500 hPa” (Sandu et al. 2013). If key near-surface processes are poorly handled in finer-scale models, their SRS parameterization in global models will be smoothed over larger grid areas and thus even less likely to be representative of the actual atmospheric processes intended to be simulated (e.g., Banta et al. 2018a). Here, we investigate error characteristics in the lowest several hundred meters of regional HRRR model versions having 3-km and 750-m grids.

### 3. Overview of WFIP2, instrumentation, and HRRR versions

**a. WFIP2: The Second Wind Forecast Improvement Project**

WFIP2 was an 18-month-long (September 2015–March 2017), multi-institutional collaborative project in central Oregon–Washington (Fig. 1a), focused on improving the quality of information and forecasts of winds and other key WE quantities within an east–west “wind-energy corridor” having many wind-generation facilities (Fig. 1b). Two major science components of WFIP2 were the 18-month field-measurement deployment extending from coastal to inland Oregon–Washington, and verification and improvement of NOAA’s HRRR operational forecast model using the WFIP2 measurement dataset. The field program consisted of arrays of ground-based in situ and remote sensing instrumentation, along with other measurement systems, as described in detail by Bianco et al. (2019: their Table 1) and in the instrumentation overview by Wilczak et al. (2019). A detailed overview of the modeling component of WFIP2 is given by Olson et al. (2019a). Datasets used in this study are publicly available from the DOE Data Archive and Portal (DAP: https://a2e.energy.gov/projects/wfip2); site addresses for individual datasets are provided with each instrument description.

**b. Instrumentation**

#### 1) DOPPLER LIDAR

Three scanning Doppler lidars were deployed along an east-northeast/west-southwest line (Figs. 1a,b) over a total distance of 71 km (Pichugina et al. 2019). Doppler lidars provide accurate wind profiles subhourly from a few meters above ground level (AGL) to 2 km AGL on most occasions. Two of the lidars (LeoSphere WindCube 200S) were installed by NOAA/CSL at sites 40 km apart at the Wasco and Arlington, Oregon airstrips. A third scanning Doppler system (Halo Streamline XR: Pearson et al. 2009) was deployed to the Boardman site, 31 km east-northeast of Arlington, by the University of Notre Dame (UND).

Table 1 shows system properties of the lidars. These instruments provided simultaneous, synchronized measurements of wind profiles using similar scanning routines and data-analysis procedures. Fifteen-minute wind profiles combined all scans within that averaging period into a velocity-azimuth-display (VAD) calculation (Fig. 3; Table 2), and then, three cycles of quality assurance produced profiles of mean direction and speed. The result is a wind profile smoothed in space.
and time, similar to smoothing inherent in NWP models (Skamarock 2004). Our use of conical scanning at low-elevation angle \(\phi\) takes advantage of smaller contamination in the horizontal-velocity calculation by the vertical-velocity term, proportional to \(\tan \phi\), in the VAD equations (see the appendix) and avoids the magnification of the radial-velocity measurement error in calculating the horizontal velocity at increasing \(\phi\) (Banta et al. 2013a). For further details on these procedures, see Pichugina et al. (2019) and references therein. Pichugina et al. (2019) also evaluated Doppler-lidar mean wind speed measurements from WFIP2 against tower data and found lidar–sonic-anemometer differences of \(0.1 \pm 0.1\) m s\(^{-1}\), consistent with Klaas et al. (2015). The 15-min profiles and associated HRRR errors were available during WFIP2 via a real-time website (Banta et al. 2021). Data from these lidars were not assimilated into the HRRR initialization, so they provide an independent assessment of model skill. Data from scanning Doppler lidars at each site are provided in A2E (2017a,b,c).

2) SW AND LW RADIATION

A site measuring the primary surface-radiation forcing for these flows was located at Wasco, in the form of a comprehensive Surface Radiation Budget Network (SURFRAD) station (see Augustine et al. 2000) measuring the surface-radiation budget [downwelling and upwelling shortwave (SW) and longwave (LW) radiation]. The station had the requisite observations for the RadFlux analysis for deriving additional clear-sky radiation variables and cloud products (Long and Ackerman 2000; Long et al. 2006; Riihimaki et al. 2019). Accuracies of NOAA’s radiation measurements range from 2% to 5% for shortwave radiation and are better than \(\pm 9\) W m\(^{-2}\) for longwave radiation (Augustine et al. 2000). Data from SURFRAD measurements are provided in A2E (2017d).

3) SEB MEASUREMENTS

Net radiation is an input to the Surface-Energy Balance (SEB); it is balanced by the surface-sensible and latent heat fluxes and the ground heat flux. Grachev et al. (2020) calculated SEB terms from the WFIP2 instrumented-tower deployment, which included sonic anemometers at 3 and 10 m at a location \(\sim 5\) km north-northeast of Wasco, referred to as the WFIP2 “physics site” (Wilczak et al. 2019). They (Grachev et al. 2020) provide a detailed description of the instrumentation deployed to this site, as well as the many challenges in measuring SEB components. Systematic and measurement uncertainty of sensible heat fluxes estimated from eddy-covariance techniques are generally \(\sim 10\)%, but the uncertainty is both positively and negatively correlated with net radiation and wind speed, respectively (Hollinger and Richardson 2005).

TABLE 1. System parameters of Doppler lidars.

<table>
<thead>
<tr>
<th>Lidar parameters</th>
<th>WindCube 200S (100-m gate)</th>
<th>Halo Streamliner XR (100-m gate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength ((\mu)m)</td>
<td>1.54</td>
<td>1.5</td>
</tr>
<tr>
<td>Pulse energy (mJ)</td>
<td>0.1</td>
<td>0.08</td>
</tr>
<tr>
<td>Pulse duration (ns)</td>
<td>400</td>
<td>350</td>
</tr>
<tr>
<td>Pulse repetition frequency (Hz)</td>
<td>10 000</td>
<td>10 000</td>
</tr>
<tr>
<td>Sampling frequency (MHz)</td>
<td>250</td>
<td>50</td>
</tr>
<tr>
<td>Accumulation time (s)</td>
<td>0.5 or 1.0</td>
<td>0.5/1</td>
</tr>
<tr>
<td>Range resolution (m)</td>
<td>100</td>
<td>48</td>
</tr>
<tr>
<td>Min range (km)</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Typical max range (km)</td>
<td>5–7</td>
<td>3–5</td>
</tr>
</tbody>
</table>

FIG. 3. Schematic depiction of scan geometry in 15-min sequence of Doppler-lidar scans, illustrating the kind of scan series referred to in Table 2. (top) Diagram of scan pattern: three nested conical (azimuth) scans indicated in blue, two vertical-slice or range-height (elevation) scans in yellow, and a representation of vertical staring in green. (bottom) Same geometry as in the top panel, but showing actual scan data on scan image; orange arrow indicates wind direction. Image from Banta et al. (2018a).

TABLE 2. Doppler-lidar scan schedule: Scan sequences continuously performed by lidars every 15 min at the three sites (Banta et al. 2021).

<table>
<thead>
<tr>
<th>Scan type</th>
<th>Elev (°)</th>
<th>Azimuth (°)</th>
<th>Duration (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wasco</td>
<td>2.75, 4, 6, 15, 45</td>
<td>0–360 scanning</td>
<td>8.5</td>
</tr>
<tr>
<td>Elevation (RHI)</td>
<td>0–30 scanning</td>
<td>0, 90, 180, 270</td>
<td>3.5</td>
</tr>
<tr>
<td>Vertical stare</td>
<td>90</td>
<td>—</td>
<td>3</td>
</tr>
<tr>
<td>Arlington</td>
<td>1.75, 3, 6, 15, 45</td>
<td>0–360 scanning</td>
<td>8.5</td>
</tr>
<tr>
<td>Elevation (RHI)</td>
<td>0–30 scanning</td>
<td>0, 90, 180, 270</td>
<td>3.5</td>
</tr>
<tr>
<td>Vertical stare</td>
<td>90</td>
<td>—</td>
<td>3</td>
</tr>
<tr>
<td>Boardman</td>
<td>1.75, 3, 6, 15, 45</td>
<td>0–360 scanning</td>
<td>6.5</td>
</tr>
<tr>
<td>Elevation (RHI)</td>
<td>0–30 scanning</td>
<td>0, 90, 180, 270</td>
<td>8.5</td>
</tr>
<tr>
<td>Vertical stare</td>
<td>90</td>
<td>—</td>
<td>3.5</td>
</tr>
</tbody>
</table>
4) MICROWAVE RADIOMETER/RASS TEMPERATURE PROFILE MEASUREMENTS

Ground-based multichannel microwave radiometers (MWRs) measure spectral downwelling radiance to provide thermodynamic-profile information above the instruments. The TROPoE algorithm (Turner and Löhner 2014, 2021; Turner and Blumberg 2019) retrieves temperature and humidity profiles from these radiance measurements using a Bayesian framework, which allows an a priori dataset to help constrain the ill-posed retrieval. The framework provides a full error characterization of the retrieved profiles, as well as profiles of information content and true vertical resolution. Recent TROPoE updates incorporate other thermodynamic data, including radio acoustic-sounding system (RASS) virtual-temperature profiles (Djalalova et al. 2022) and multielevation angle MWR measurements (Turner and Löhner 2021).

The present study used MWR brightness temperatures at eight frequencies along the 22.2-GHz water-vapor line and 14 frequencies along the 60-GHz oxygen-absorption band at elevation angles of 19.8°, 90°, and 160.2°. These instantaneous brightness temperatures were combined with hourly RASS virtual-temperature profiles, linearly interpolated in time, and near-surface temperature and humidity measurements, into 15-min TROPoE retrievals. Djalalova et al. (2022, Fig. 8b) found measurement RMSEs of 1–1.4 K, for temperature retrievals averaged over 0 to 3 km AGL, using TROPoE bias correction. Data from microwave radiometers and RASS are provided in A2E (2017e,f,g).

c. Experimental HRRR versions

Here, we use NOAA’s High-Resolution Rapid Refresh system (Olson et al. 2019a; Dowell et al. 2022), a short-term forecast model run hourly 24/7 at NCEP to provide forecasts issued to the public. It was developed at NOAA/GSL, which is responsible for biennial improvements and updates. This 3-km-grid model, nested in the 12-km Rapid Refresh (RAP) model (Benjamin et al. 2016), is widely used by aviation, agriculture, renewable energy, severe weather, and others, because of its ability to assimilate the latest atmospheric measurement data hourly. Both models used the Advanced Research version of WRF (WRF-ARW) as the underlying framework.

Early WFIP2 HRRR evaluations of wind profiles (Pichugina et al. 2019) used scanning Doppler-lidar data and output fields from NCEP’s operational HRRR at the time, HRRR-NCEP version 1, which ran on the full HRRR grid (orange box in Fig. 1c) for each hour of the 18-month WFIP2 campaign. Using this full 24/7/365 operational HRRR dataset enabled the largest possible sample sizes and representative annual error statistics.

In developing new versions, 3-km-grid experimental test runs were performed on a smaller “Provisional WFIP2-HRRRNEST” domain (small green box), for more limited time periods, to conserve limited computing resources. The present study uses test versions 1 and 4 (v1 and v4), and their nests, as indicated in the more limited green-box domains. The v4 summer tests were only run for the August period. HRRR’s terrain representation at 3 km and 750 m within the study region is shown in Figs. 1d and 1e. The more filtered 3-km terrain smooths many terrain obstacles that would retard the flow in 750-m runs and in the atmosphere.

Updates to each version of HRRR from v1 to v4 are tabulated by Dowell et al. (2022). Of relevance here are an eddy-diffusion mass-flux (EDMF) boundary layer diffusion scheme, including shallow subresolution clouds (starting in version 2), soil-type and albedo modifications allowing MODIS satellite input, updates to the MYNN-ABL schemes as they became available, and orographic drag. Adler et al. (2023) describe updates to v4 that specifically address stable mixing, including reducing the stable-mixing-length scale, calculating horizontal diffusion along geopotential Cartesian levels rather than terrain-following, reducing the diffusion parameter in the sixth-order filter (added in v2) to minimize steep-terrain filtering, and updating the gravity wave-drift scheme. In their HRRR evaluation of a wintertime cold-pool episode, Adler et al. (2023) concluded that these v4 updates reduced vertical diffusion to more representative atmospheric values, consistent with previous findings (Olson et al. 2019b; Berg et al. 2021).

Stable-mixing issues are complicated by numerical (“implicit”) diffusion (e.g., Smolarkiewicz 1982, 1983) and imposed minima on the allowable values of the SRS diffusivity, which may appear as specified minima of TKE or other turbulence variables (set generally at model-run initialization). A smoothing filter may also be applied. Altogether, they enforce a minimum or floor on the allowable magnitude of diffusion in the models—a baseline diffusion value at all model points—to avoid numerical instability and spurious wave activity. For stable conditions, this baseline minimum is often greater than corresponding atmospheric values, leading to excessive mixing in the models (e.g., Sandu et al. 2013).

Studies have evaluated HRRR wind errors from instrumented-tower data (Olson et al. 2019a; Lee et al. 2019; Fovell and Gallagher 2020; Turner et al. 2020; James et al. 2022). For WE and NWP improvement, the accuracy of modeled winds aloft is also critical. Benjamin et al. (2016) and Fovell and Gallagher (2020) used rawinsonde-profile data in validations of RAP and HRRR. Banta et al. (2014, 2018a), Djalalova et al. (2016), and Pichugina et al. (2017) evaluated HRRR against an offshore dataset of remotely sensed, ship-borne lidar wind measurements over the Gulf of Maine. As already alluded to, several studies have used WFIP2 ground-based remote-sensor datasets to evaluate HRRR versions (Pichugina et al. 2019, 2020, 2022; Olson et al. 2019a,b; Bianco et al. 2019; Banta et al. 2020, 2021; Ghate et al. 2022; Adler et al. 2023), and other studies have simulated WFIP2 cases (Berg et al. 2021; Liu et al. 2022). As summarized by Banta et al. (2021), “these studies taken together indicate representative HRRR RMSE values of approximately 3 m s⁻¹ and bias magnitudes of generally 0.5–1.0 m s⁻¹ overall.”
4. Results

a. Error evaluation by Doppler lidar

We first use wind speed data from scanning Doppler lidars at Wasco, Arlington, and Boardman (Fig. 1) to evaluate errors in these models for the marine-intrusion days. Figure 4 shows time–height cross sections of lidar-measured (top row) and HRRR-modeled (next two rows) windspeeds, for the three lidar sites on an individual marine-intrusion day. The bottom two rows show the model–measurement differences (errors) for v1 and v4.

Fig. 4. Time–height cross sections of wind speed (color bar; m s⁻¹) for an individual marine-intrusion day, 15 Aug, at (left) Wasco, (center) Arlington, and (right) Boardman. (a) Doppler-lidar-measured wind speed, (b) model cross sections for v1, and (c) model cross sections for v4 [arrows in (a)–(c) show wind direction, and dark bars along time axis of row three represent nighttime hours (0300–1300 UTC)]. Model-minus-measurement differences (model errors) for (d) v1 differences and (e) v4 differences.
This example illustrates how both versions ended the strong flow below 600 m AGL too soon (before 0800 UTC), at Arlington and Boardman, whereas lidar-measured winds of >8 m s\(^{-1}\) persisted until after 1400 UTC. In the HRRR predictions, the wind speed drop was accompanied by a transitory shift from westerly to northerly or northwesterly flow [see also Banta et al. (2020), their Fig. 9 and Pichugina et al. (2019), their Fig. 10]. Lidar-measured winds do not show this shift at Arlington below 400 m, and although the lidar winds at Boardman here show a brief deflection of the winds to WNW between 100 and 400 m AGL (but not below); 0500 UTC, this shift to northwesterly flow was not seen in the measurements on any other intrusion days below 400 m. The model–measurement differences show that the premature termination of the strong winds produced errors of more than 4 m s\(^{-1}\) by both models, most strongly at Arlington and Boardman, and that Wasco also saw large errors, but aloft above 200 m AGL (also as previously found).

Composite cross sections (Fig. 5) for the four August intrusion days show the premature end of the strong flow and the associated large low biases (negative mean errors of greater than 4 m s\(^{-1}\)) in both versions of HRRR, as also seen on each individual intrusion day. Analyses are shown for each site, and for the three-site average; the last, shown to correlate more strongly with wind-power generation over the region, by smoothing out the temporal discrepancies due to the advance of wind-change phenomena (such as marine-intrusion fronts) through the sites (Pichugina et al. 2020, p. 7). The strength of the initial wind surge at 0300–0400 UTC was captured by all versions at Wasco but was underpredicted at Arlington and Boardman. A significant, erroneous layer of strong winds aloft between 0600 and 1200 UTC at Boardman in the models accompanied the shift to northwesterly flow at that location. Associated with a growing nocturnal stable-boundary layer (SBL) inversion, the measurements show a surface-based layer of weak winds growing in depth through
the night, most apparent after 0700 UTC at Wasco. This deepening layer of weak winds was not well captured by the models, e.g., Wasco’s bias plots show large overpredictions of wind speed below 200 m. Additionally, the large underpredictions due to the intrusion demise in the lowest ~500 m at Arlington and Boardman were significantly moderated below 100 m, to less than 1 m s\(^{-1}\). At Arlington, this shows up in the v1 bias plot as a thin ground-based layer of smaller values of absolute bias (or |bias|: lighter-blue-shaded layer) below 100 m, also seen in v4 but less strongly. Profile analyses presented below will further clarify these findings.

Composite cross sections of the 750-m, nested configurations of each version (Fig. 6) indicate smaller bias magnitudes than their parent versions overall. Wind speeds still decreased wrongly after 0800 UTC, but the underestimates were smaller than for the parent versions. Most notably, v1-nest wind-speeds below 100 m at Arlington and Boardman were stronger and thus closer to measured values than those of the other versions, and thus, the v1-nest produced a significantly smaller absolute-bias than the parent versions below 200 m AGL (less than 0.5 m s\(^{-1}\)) between 0600 and 1600 UTC at these sites.

Comparing v4 and v1, the upper rows of each pair in Fig. 7 show the changes in magnitude of the bias (i.e., changes in “absolute-bias”) \(\Delta|\text{bias}|\) for the parent versions resulting from updates to model physics and other schemes from v1 to v4, which we will simply call physics updates. Negative values, indicating error reductions and model improvement for v4, appear at night (0800–1400 UTC) near the surface at Wasco and above 150 m at the other two sites. The updated v4 produced larger errors, however, below 150 m at Arlington and Boardman at similar times. This interesting result is consistent with the previous WFIP2 findings. Changes in RMSE (\(\Delta\text{RMSE}\), Fig. 7b), exceeding 2 m s\(^{-1}\) increases for v4 at Arlington-Boardman below 150 m, showed a similar overall pattern to \(\Delta|\text{bias}|\).

The skill improvements due to nesting are illustrated in Fig. 8a, which shows \(\Delta|\text{bias}|\) due to using finer grid resolution. Reductions in error were most dramatic for v1, exceeding 2 m s\(^{-1}\), and smaller-magnitude improvements are also seen
for the v4 nesting. ΔRMSE (Fig. 8b), exceeding 2 m s⁻¹ reductions for v1 and smaller reductions for v4, again showed a similar overall pattern to Δ|bias|.

1) VERTICAL PROFILES

Figure 9 shows vertical profiles of the mean wind and associated errors averaged over the nighttime 0300–1300 UTC period, and Fig. 10 shows the lowest 200 m on an expanded vertical scale. Mean-wind profiles show observed and simulated low-level-jet (LLJ) structure at all sites, the modeled jet maxima generally weaker and lower than observed. The simulated jet for v1 at Wasco is an interesting case. Its flat 100–300-m LLJ profile was 3 m s⁻¹ weaker than the 12 m s⁻¹ measured jet maximum at 300 m, but the v1 winds below 160 m were stronger than the corresponding observed winds at the same level, leading to errors of >2 m s⁻¹ below 100 m, the largest error appearing as a sharp spike in the profile at 40 m AGL. The reason for this errant LLJ structure at Wasco is unknown, but it does not appear in the other HRRR versions, such that v4 and the nests all show error reductions (ΔRMSE column) of 2 m s⁻¹ at 40 m at Wasco compared with the v1-parent.

At Arlington and Boardman, the dramatic premature drop in wind speed noted in Fig. 4 dominates the error profiles, resulting in parent-version underpredictions of 3–4 m s⁻¹ and RMSE’s exceeding 4 m s⁻¹ in layers up through at least 400 m AGL. However, as at Wasco, the v1 profiles at these sites show flat shapes below 100 m with a sharp bend just below 40 m (see Fig. 10). The v1 winds were not as strong as observed, but they were stronger than v4’s, such that the near-surface winds in v1 were in better agreement with the measured winds at these sites than the v4-parent. Thus, a reason that the v4 wind forecast was worse than v1 as seen in Fig. 4 is that v1 has generated a false low-level wind maximum below 100 m that v4 has corrected for. But because the winds were systematically too weak at these sites to begin with, due to the premature intrusion demise, the erroneously stronger winds of v1 are actually in better agreement with the measurements, by an error-cancellation effect.

The modeled mean-wind profiles were smoother through their depth and had weaker shear than observed (especially evident in the v1 profile at Arlington), which can be considered diagnostic of the excessive model diffusion discussed in section 3c. The v1 winds were weaker than v4’s from 150 to 400 m, but below 150 m, the v1 wind speeds were stronger. Model diffusion mixed stronger flow above 150 m down to these near-surface levels, to produce a tendency toward higher wind speed bias there. This mixing appears less effective in the v4 profile, due to its modifications to the stable-mixing scheme.

We have identified three factors contributing to the wind speed errors below 400 m at Arlington and Boardman. The simulated marine-intrusion demise produced large and obvious underpredictions. Then, v1 generated an errant LLJ in this layer below 100 m that compensated for this large error,
especially at Wasco. Finally, below 100 m, the large intrusion-demise error was offset by positive error contributions from excessive model diffusion. Version 4 had weaker, more realistic diffusion and did not generate the false near-surface LLJ. Thus, ironically, v4’s more realistic schemes to reduce errors were less effective in offsetting the large negative errors, with a result that v4’s skill was worse than v1’s in predicting wind speeds below 150 m. This seems to be a situation where, roughly stated, better physics produced worse forecasts. Caution in interpreting model error increases or decreases as corresponding changes in model skill is warranted; chasing down the various contributions to model error is still worthwhile, however, to better understand how to approach improving the models.

Another noteworthy aspect of the profiles is the significant reductions in model error achieved by finer grid resolution. The smallest errors of all below 200 m at these sites were achieved by the v1-nest, and in general, the nested versions outperformed the parent versions through surface-based layers several hundred meters deep.

The right two columns of Figs. 9 and 10 show the vertical structure of the changes in error among versions. Whereas the bias magnitude and RMSE below 200 m (Fig. 10) were mostly 3 m s⁻¹ or more at Arlington and Boardman, and thus should be discernable by most vertically profiling remote wind sensors, the error differences in the right columns are smaller, mostly less than 1 m s⁻¹, which would require accurate measurements to properly quantify. Because Doppler lidar provided the reference measurements for these error calculations, the reported Doppler-lidar uncertainties of 0.1 m s⁻¹ vouch for the reliability of these version-difference values.

The substantial error reductions, due to nesting v1 as compared with the v1-parent (yellow profiles), amounted to more than 1 m s⁻¹ in absolute-bias and RMSE through at least a 600-m deep layer at Arlington and Boardman. Error-change profiles at Arlington and Boardman for the v4-parent compared with v1 (light blue) above 150 m show the depth, structure, and magnitude of the error reductions there due to the physics updates. Below 150 m as just discussed, the significant decreases in errors at Wasco were due to the more realistic LLJ structure in v4, and the error increases for v4 at Arlington and Boardman below 100 m are seen to peak at 1 m s⁻¹.

2) TIME SERIES

Time series of the wind speed and errors averaged vertically over 50–150 m (Fig. 11) show how all versions underpredicted the winds at Arlington and Boardman after 0300 UTC, the v1-nest performing the best and the v4-parent, the worst overall. The smaller errors in the v1-nest than those of the other versions are evident at Arlington and Boardman between 0600 and 1500 UTC. The LLJ-induced overpredictions of speed by the v1-parent at Wasco grew in time to produce an upward ramping of RMSE from 0800 to 1400 UTC.

Fig. 8. Time–height cross sections of error differences as in Fig. 7, but between versions due to reducing grid interval (nest minus parent). (a) Δ|bias|, the difference in bias between versions for v1 in the first row and v4 in the second row. (b) Δ-RMSE, the difference in RMSE between versions.
Fig. 9. Vertical profiles of lidar-measured (black curve) and (first column) modeled wind speed and model-error statistics averaged over the nighttime hours of large error (0300–1300 UTC) for each site and for the three-site mean. The error statistics are shown for (second column) bias and (third column) RMSE. Version-to-version differences (v4 minus v1 or nest-minus-parent) in the error statistics are shown for (fourth column) |bias| and (fifth column) RMSE. Vertical axis: height AGL up to 1 km; horizontal axis: wind speed or speed difference (m s⁻¹).
reaching more than 3 m s$^{-1}$, which the nests and v4 mostly corrected for.

The error-difference plots for the physics updates in the parent versions (Fig. 12a) show the decreases in error for v4 at Wasco, and that the increased v4 errors at Arlington occurred during 0300–0700 UTC, when stronger flow mixed down into this layer in v1 than v4 (Fig. 4) as described, to offset the large underpredictions. At Boardman, the largest v4-error increases over v1 in this layer occurred after 1000 UTC. The physics updates to the nested version (Fig. 12b) produced degradation of the solution at Arlington and Boardman from 1000 UTC until daytime heating became strong, but otherwise mostly neutral results. The significant error reductions for v1 due to nesting can be seen in Fig. 12c, whereas improvements from nesting v4 (Fig. 12d) exceeded 1 m s$^{-1}$ only at Arlington. Comparing v1-parent with the v4-nest produced some improvement at Wasco and Arlington, but overall less than the improvement between v1 and its nest. During the noteworthy time periods discussed, the magnitude of the model differences significantly exceeded the accuracy threshold of the lidars.

3) SUMMARY PLOTS

Figure 13 summarizes the mean error (upper panels of each pair) and the change-in-error (lower panels) results for absolute-bias and RMSE for the 50–150-m layer, averaged over 24 h (left-column panels) and over the 0300–1300 UTC nighttime periods.
The top panel for each error type shows the average error for v1 (red) and v4 (blue) (nested versions cross-hatched). RSMEs generally hover near 3 m s\(^{-1}\), as in previous studies, and nested versions, mostly smaller.

The lower (second and fourth rows) panels summarize the changes in error due to the physics and grid updates. As seen for the profile measurements, the magnitudes of the version-to-version differences were smaller than the errors, typically

![FIG. 11. Time series of lidar-measured (black curve) and (top) modeled wind speed calculated at 15-min intervals for the 50–150 m AGL layer, representing the rotor layer of a hypothetical wind turbine. The corresponding time series for (middle) bias and (bottom) RMSE for each HRRR version.](image)

![FIG. 12. Time series as in Fig. 11, but of error differences (Δ-error) between model versions. (a) Physics updates, parent versions (v4 minus v1); (b) physics updates to nested versions (v4-nest minus v1-nest); (c) finer grid, v1 (v1-nest minus v1); (d) finer grid, v4 (v4-nest minus v4); and (e) combined physics plus nesting updates (v4-nest minus v1). The thick solid line indicates Δ|bias|, and the thin line with dots shows ΔRMSE. The vertical axis in all plots is in m s\(^{-1}\).](image)
FIG. 13. Error summary plots for (a),(b) |bias| and (c),(d) RMSE, averaged over the 50–150-m layer for (left) the entire day (0000–2400 UTC) and (right) the nighttime periods (0300–1300 UTC). Top plots in each row pair show the magnitude of each error statistic for each site and the three-site mean, color-coded by version (v1, red; v4, blue; nests shaded). Bottom plots in each pair show error differences, color-coded by which versions are being compared.
less than 1 m s$^{-1}$. The significant reductions in absolute-bias and RMSE at Arlington and Boardman due to nesting of v1 (yellow) are clear, but advantages of nesting v4 (orange) are smaller to nil in this layer, except at Arlington. Parent-version physics updates in v4 (light blue) produced error decreases at Wasco due to the improved representation of the LLJ structure at night in v4. At Arlington and Boardman, the v4 updates mostly increased the bias magnitudes, as the v4 LLJ and stable-mixing scheme changes described above generated absolute-bias increases of more than 0.4 m s$^{-1}$, well above the lidar accuracy limits, and produced RMSE increases at Arlington but not Boardman. In the nested versions (darker blue), v4 physics updates made little change at Wasco but degraded the results at the other sites, the largest error increases indicated for the nighttime period.

Overall, the physics and numerical updates to v1 that produced v4 did not produce consistent error reduction in the wind speed forecasts for this type of wind flow across all sites and versions for various reasons. Thus, interpretation of these results is complicated, but one result of note is that reducing the grid interval did produce smaller absolute-bias and RMSE. The more faithful representation of the topography at 750 m (Figs. 1d,e) allows for more realistic channeling of the flows and a more realistic representation of the horizontal patterns of surface-heating variability; the more rugged terrain obstacles would more effectively impede the flow; and a range of small-scale flows that would be modeled as SRS diffusion at 3 km would be modeled as transport at 750 m, which is often unrelated to local 3-km resolved-scale gradients (Banta et al. 2018a).

b. Discussion: Doppler-lidar results

The error patterns varied significantly in space (site to site and vertically) and time, but three effects appeared dominant. First, diffusive mixing of momentum downward through the lowest several hundred meters of the nighttime SBL was a key contributor. Second, unknown model processes produced an erroneous jet peaking at 50–100 m in v1. Third, the premature weakening of the winds dominated the flow below 200 m at Arlington and Boardman. It was the interplay among these effects—not necessarily which version had more realistic physics schemes—that determined which model version produced winds that agreed more closely with the measurements.

Still unresolved is why the HRRR winds died off too soon at Arlington and Boardman. The weakening of the flow in the models was accompanied by a windshift to northerly or northwesterly on intrusion nights. Animations of the HRRR 80-m winds (see supplemental material) show that another pulse of strong marine-intrusion outflow originating from farther north through Snoqualmie Pass pushed southeasterly and southerly toward the Columbia River Valley near Arlington and Boardman. Images from the animation (Fig. 14) show the flow from this gap spreading southeastward at 0100 UTC, and northerly winds pushing all the way to the Columbia Valley by 0400 UTC, where speeds reached 14 m s$^{-1}$. Vertical north–south cross sections through Arlington and Boardman (supplemental material) show that the enhanced wind speeds immediately north of these sites were due to a hydraulic jump-like flow (a lee wave or modest “windstorm”) generated by the model over the ridge to the north, producing strong northwesterlies along the lee slopes of this ridge but weak flow over the measurement-site locations in the Columbia Valley.

Important questions become, did the real outflow from the Snoqualmie Pass area push northerly flow to the vicinity of Arlington and Boardman? and did the atmosphere generate a mountain wave there, but not strong enough to penetrate to the measurement sites? The first question is important because if the model is generating a too-vigorous sea-breeze circulation, then some aspect of model physics is likely to blame and could be diagnosed and potentially fixed, but if the errors were due to the misrepresentation of the terrain in the model, then updating the model physics would not fix the problem. The second question is important because if the model is producing erroneous values of shear and lapse rate, such that mountain waves are being generated where they did not occur (or failing to predict them when they do occur), this would be detrimental to the forecasting of low-level wind speeds. This would be important information for WE forecasters to be aware of, for example.

Lidar time–height cross sections at Arlington and Boardman did show some evidence of northwesterly winds aloft on some marine-intrusion days above 300 m. Whether these could be related to the effects just discussed is an important but unanswered question.

5. SEB and thermodynamic verifications

Primary sea-breeze forcing is the difference in daytime surface heating between the land and sea. Modeling of such flows requires accurate representation of each of the “links of the chain of processes that need to be properly modeled” (Banta et al. 2021). These processes include interception of solar radiation at the surface, heating of the near-surface air via SEB interactions, and generation of the cross-coastal pressure gradient, which drives the landward acceleration of the wind. A hypothesis for the anomalous northerly flow reaching Arlington and Boardman in v1 is that an overprediction of heat fluxes inland would produce an overactive sea-breeze circulation, which would push the outflow from Snoqualmie pass too far inland, resulting in the northwesterly flow over the ridge north of these two sites. Here, we evaluate HRRR’s ability to simulate the surface-radiation budget, the SEB, and the near-surface temperature over land, where the diurnal variation is much stronger than over the ocean, where measurements are limited.

a. Radiation budget and surface-energy balance

Figure 15 shows the measured and modeled downwelling shortwave and longwave radiative flux (incoming solar radiation) at Wasco for the 4 days. The skies on 15 and 17 August were clear. On 16 August and the morning of 14 August, patches of optically thin cirrus passed over, as indicated by transient dips in the measured downwelling shortwave and increases in the measured downwelling longwave fluxes. The
similarity of the evolution of the intrusion on all four of these days suggests an insensitivity of these flows both to the minor variations seen in the downwelling shortwave and to the occurrence of thin cirrus layers. Model errors of the downwelling radiation were small and near measurement thresholds of the instruments, except during cirrus occurrences. Although v4 appears to have improved the SW-down midday on 17 August, the magnitudes of the model version...
differences for these two components were less than the measurement accuracy, except for the cirrus period early on 14 August.

The four radiation budget terms and net radiation, the errors for each model version, and the change in absolute error from v1 to v4, for the clearest of the 4 days, 17 August, are shown in Fig. 16, as measured at Wasco and modeled by v1 and v4. Downwelling daytime solar fluxes were modeled mostly to within 15 W m$^{-2}$ of the measured values, the version differences agreeing to within the measurement uncertainty, as noted. The reflected, upwelling midday shortwave was significantly underestimated (40 W m$^{-2}$ or $\sim$25%) by v1, but the surface albedo updates implemented to correct for this in v4 overestimated this component by a smaller but still significant amount (25 W m$^{-2}$ or 17%). The raw version differences reach more than 70 W m$^{-2}$, significantly greater than even the 5% measurement-error threshold. The largest errors to the longwave components reached 20 W m$^{-2}$ at night, with smaller values during daytime, but the version differences were mostly within the measurement error of 9 W m$^{-2}$. Errors in the net radiation ($R_{net}$, right column) were dominated by the large upwelling shortwave errors, indicating the surface albedo issues in v1 being overcompensated by v4. Midday v1 values of $R_{net}$ were too large by 20–30 W m$^{-2}$, whereas the v4 values were too small by a similar amount.

Net radiation is an input to the SEB as shown in Fig. 17, which also includes the measured and modeled sensible heat flux $H$ for the 4 days and modeled values for other SEB components (due to missing/incomplete observation records). The differences in midday $H$ values between v1 and v4 were 35–55 W m$^{-2}$, similar to but somewhat smaller than the differences in $R_{net}$ (Fig. 16j). Both model versions overestimated midday $H$, the overprediction in v1 being nearly twice that of v4 (Fig. 17b).
b. Mean low-level temperature difference

Figure 18 shows daytime temperature-bias distributions at Wasco, averaged vertically over 50–200 m, for each model version, using the microwave radiometer/RASS data combined in the TROPoP calculation. All HRRR versions overestimated the temperature, v1 having the largest bias and v4-nest having the smallest, this difference amounting to 0.37 °C. The positive bias and the fact that it was largest for v1 are consistent with the noted overpredictions of $R_{\text{net}}$ and $H$ and with the hypothesis of an overactive sea-breeze circulation. The differences are less than the measurement accuracy, however, so more accurate temperature-profile measurement would be desirable for model-improvement work. Warmer inland temperatures lead to lower pressures over land, such that model versions generating the warmer temperatures should produce larger coastal–inland pressure differences and drive a stronger sea breeze. Comparing model and measured pressure differences involves many complications, including pressure height adjustments, model smoothing, and spatial representativeness of pressure values. We attempted to demonstrate this effect, but the saved surface-pressure output variables differed between v1 and v4 and were incompatible for comparing pressure differences among the versions.

In summary, $R_{\text{net}}$ was overestimated by v1 and underestimated by v4 due to differences in albedo, but both HRRR versions overestimated $H$, indicating that unknown SEB terms, most likely ground heat flux in this arid landscape, were mainly responsible. But $R_{\text{net}}$ was still a factor, as v1, which had the largest $R_{\text{net}}$, had an error value of $H$ twice as large as v4, and the difference in magnitude in $H$ between the two versions was similar to the difference in $R_{\text{net}}$ (~50 W m$^{-2}$). The larger heat fluxes in v1 led to higher mean low-level temperatures than v4 by 0.25 °C, which would be expected to produce lower inland pressure than experienced by the atmosphere, which in turn would drive a more active sea-breeze circulation in the model. These results are consistent with the overactive sea-breeze hypothesis, but difficulties in measuring the SEB and the noisiness of the heat-flux measurements preclude solid confirmation. Determining what the relevant area-averaging for heat flux is, and what is the best way to measure it, are needed areas for research.

6. Conclusions

These WFIP2 marine-intrusion NWP evaluation studies have shown that the WRF-derived HRRR model generated large wind speed errors when simulating a flow driven by purely dry processes. It did so repeatedly, in the same manner, each time the atmosphere produced this recurrent flow system, the marine intrusion.

Here, we have found that the largest errors could be associated with an active outflow in the model through a pass farther north outside the study area. Although the precise cause
of these errors could not be determined with this dataset, we offered some potentially contributing hypotheses. The deeper penetration of the flow into the CRV could result from over-active sea-breeze forcing mechanisms; flow channeling by the model topography may have been wrongly represented, especially for the coarser 3-km grid; and the smoothed topography may have insufficiently retarded this flow.

Errors are complex, even for this seemingly straightforward sea-breeze-type flow that is largely isolated from larger-scale influences. What we have been able to do with the WFIP2 dataset was to demonstrate this complexity of model errors and to compare the error behavior among model versions at the available sites. HRRR itself, as other NWP models, is complex, and v4 differs from v1 by a large number of updates, as listed by Dowell et al. (2022). It was no surprise that the errors and error differences were also complex.

One kind of complexity was the spatial variability of the errors. The nature of the model errors differed among the three measurement sites, as did model-version skill changes. These significant site-to-site differences in HRRR errors could be interpreted as indicating error randomness, but at each measurement site, the diurnal error signature was the same from case to case for all marine-intrusion days. HRRR thus replicated the same horizontal pattern of the error systematically whenever this type of flow occurred. The error pattern itself was complex and most likely not well sampled by three wind-profile sensors. Research planning should address this kind of variability.

The generation of unrealistic flows or vertical structure by the model can complicate error evaluation. The reason for the large deviation of the wind profile at Wasco in v1 below 200 m is
unknown but did not appear in v4 or the nests, so all the other versions showed improvement in this layer compared with v1.

Complexity was also evident in the compounding of model errors. We found that by far the largest errors were the premature demise of the strong marine-intrusion flow at two sites. These large errors were lessened in the lowest 200 m by excessively strong near-surface model diffusion bringing higher-momentum air downward to produce stronger winds more in line with the measurements via error compensation. It was thus ironic that the v4 diffusion, less excessive and thus more realistic, resulted in larger model errors than v1, because of its being less effective in compensating for the large underprediction errors of the intrusion demise.

Doppler-lidar profiles allowed us to see the effects of the excessive model diffusion, where model profiles were significantly smoother and had weaker shear than observed especially below 500 m. As described, excessive diffusion is an inherent property of finite-difference models, most strongly affecting stable conditions. Because large-scale flow properties such as 500-hPa wave patterns and statistics in global models have been shown to be sensitive to stable-mixing properties near the surface, it is likely that the existence of this baseline diffusion in the models would be an ultimate limitation on attainable model skill for a given grid resolution. It would be useful to determine that limit by measurements or other analysis because as models approach the limit, further significant improvement will not be achievable. Further skill enhancement would require finer grid resolution.

A clear result was that nesting improved HRRR skill for the marine-intrusion problem more than the physics updates, consistent with WFIP2 seasonal and annual error determinations over all types of flow in previous WFIP2 studies. This has not been a universal finding, as several studies (mostly verifying precipitation) have found that finer resolution failed to enhance skill (e.g., Mass et al. 2002; Mittermaier 2014), and Banta et al. (2018a) showed results from an offshore dataset where the finer-resolution model (HRRR) increased the errors for lead times of more than 4 h as compared with the parent RAP results. On the contrary, WFIP2 modeling studies have consistently shown, for all averaging periods, that the finer grid resolution of HRRR-nests produced the best forecasts, as also confirmed for the dry-physics cases studied here.

The v1-nest had the best performance of all versions. Nesting of v4 had a less significant effect, reducing RMSEs below 200 m only at Arlington. That the v4-nest did not improve on the v1-nest could mean many things, such as, it could mean that the adjustable “constants” in v4 need to be retuned, or that parameterizations in v1 are better suited to 750-m grids than those of v4, or that parameterizations and other model aspects need to be completely reevaluated for running on finer grids.

In several respects, it was clear that the WFIP2 dataset lacked adequate measurements to resolve some key modeling issues. Any wind profiles to the north of Arlington and Boardman would have answered whether the northwesterly flow from Snoqualmie Pass penetrated to this ridge in the atmosphere as it did in the model, addressing the degree to which the model may have generated an overactive sea breeze. More accurate SEB measurements, distributed over a wider area, as well as a sense of what the appropriate area-averaged value of the surface heat flux is and how should that be measured, would have clarified the role of this flux in generating errors in the primary sea-breeze forcing. More accurate temperature profiles to accompany the wind profiles would test a new model version’s ability to better reproduce the thermodynamic and stability fields. Increasing the number of profiling sites anchored by scanning Doppler lidars by a factor of 3 or more over an area such as WFIP2 would more adequately sample the horizontal variability of the winds, the model errors, and the model-error differences among versions. These are some of the many issues that meteorological analysts, instrument specialists, and NWP modelers intent on using the resulting datasets should jointly consider in designing field campaigns for model improvement. Such datasets would be invaluable in the development and validation of artificial intelligence forecasting algorithms.

A basic step in the model-improvement process is determining which modeled processes are responsible for the largest errors. Here, guidance from measurement campaigns is needed: comprehensive, holistic campaigns similar to WFIP2 in that all aspects of the flow, its forcing mechanisms, and its boundaries are measured. The goal of such campaigns should be “to specify as completely as possible the three-dimensional state of the atmosphere for successive time periods (at perhaps 10- or 20-min intervals)” (Banta et al. 2013a,b) and thereby to characterize as fully as current technology and resources allow, what it is that the model is trying to replicate. Analyses of such a dataset would be a target for models to aim at, comparing model versus measured 3D meteorological fields as they evolve.

Acknowledgments. The authors thank WFIP2-experiment colleagues Scott Sandberg (CSL), Clark King (PSL), Ann Weickmann (CSL), and Aditya Choukulkar for preparation and deployment of lidars to the research sites. From NOAA/ESRL, we thank Joe Olson (GSL), Roy Miller (CSL), Tilden Meyers (GML), James Wilczak (PSL), and Melinda Marquis (GSL). We also thank three anonymous reviewers for careful reading and helpful reviews of this manuscript and Editor J. R. Minder for a fourth comprehensive and insightful review. This work was sponsored by the NOAA/CSC Air Quality Program, the Atmospheric Science for Renewable Energy (NOAA/ASRE) Program, in part by the NOAA Cooperative Agreement with CIRES, A170AR4320101, and NOAA was funded in part by the U.S. Department of Energy, Wind Energy Technologies Office, via DOE Grant DE-EE0007605; Vaisala, Inc. (M.T. Stoeving) was funded by DOE Contract DE-EE0006898; Notre Dame lidar deployment and Sharply Focused, by Subcontracts DOE-WFIFP2-SUB-001 and DOE-WFIFP2-SUB-003.

Data availability statement. WFIP2-sensor datasets and output for the HRRR experimental model runs are available on the DOE Data-Archive and Portal (DAP): (https://a2e.energy.gov/data#wfip2); individual datasets are listed with the references. The HRRR code is included in WRF repository, and the information needed to replicate the (physics) configurations is listed in Dowell et al. (2022, their Table 4).
TABLE A1. Reduced 2D along-wind VAD: variables and values. Here, M = measured, specified; U = unmeasured: calculated from measured quantities.

<table>
<thead>
<tr>
<th>M/U</th>
<th>Upstream Value</th>
<th>Downstream Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured radial velocity component</td>
<td>VR</td>
<td>-VR</td>
</tr>
<tr>
<td>Scan azimuth (θ), cosθ</td>
<td>θw cosθu</td>
<td>180°, -1</td>
</tr>
<tr>
<td>Horizontal wind component</td>
<td>Uu</td>
<td>Ub</td>
</tr>
<tr>
<td>Vertical wind component</td>
<td>wu</td>
<td>w</td>
</tr>
</tbody>
</table>

APPENDIX

2D, Along-Wind Reduced VAD

Solve the radial-wind VR equation VR = U cosθ cosh + w sinh (Browning and Wexler 1968) for the horizontal wind component U:

\[ U_e = \frac{VR}{\cos \theta \cos \phi} - \frac{w \tan \phi}{\cos \theta_e} , \]

where \( \phi \) = elevation angle, \( e = u \) or \( d \) (upwind or downwind), and Table A1 defines other symbols. Adding upwind and downwind components and substituting from Table A1 yields the following:

\[ U_h = \frac{1}{2} \left[ \frac{VR_d - VR_u}{\cos \phi} + (1 - \xi) w \tan \phi \right] . \]

To maintain generality, we have set \( w_u \) = its upstream value \( w \), and \( w_d \) as proportional to \( w \), via a vertical-velocity differential factor \( \xi \). The resulting \( w \)-adjustment term depends on \( \phi \), except when \( w = 0 \), or for the special case where \( w \) is proportional to \( \cot \phi \). See the online supplemental material for further derivation details.

REFERENCES


