Downslope Windstorm Forecasting: Easier with a Critical Level, but Still Challenging for High-Resolution Ensembles

JOHNATHAN J. METZ[a] AND DALE R. DURRAN[a]

[a] *University of Washington, Seattle, Washington*

(Manuscript received 22 July 2022, in final form 8 April 2023, accepted 24 May 2023)

ABSTRACT: Strong downslope windstorms can cause extensive property damage and extreme wildfire spread, so their accurate prediction is important. Although some early studies suggested high predictability for downslope windstorms, more recent analyses have found limited predictability for such winds. Nevertheless, there is a theoretical basis for expecting higher downslope wind predictability in cases with a mean-state critical level, and this is supported by one previous effort to forecast actual events. To more thoroughly investigate downslope windstorm predictability, we compare archived simulations from the NCAR ensemble, a 10-member mesoscale ensemble run at 3-km horizontal grid spacing over the entire contiguous United States, to observed events at 15 stations in the western United States susceptible to strong downslope winds. We assess predictability in three contexts: the average ensemble spread, which provides an estimate of potential predictability; a forecast evaluation based upon binary-decision criteria, which is representative of operational hazard warnings; and a probabilistic forecast evaluation using the continuous ranked probability score (CRPS), which is a measure of an ensemble's ability to generate the proper probability distribution for the events under consideration. We do find better predictive skill for the mean-state critical-level regime in comparison to other downslope windstorm–generating mechanisms. Our downslope windstorm warning performance, calculated using binary-decision criteria from the bias-corrected ensemble forecasts, performed slightly worse for no-critical-level events, and slightly better for critical-level events, than National Weather Service high-wind warnings aggregated over all types of high-wind events throughout the United States and annually averaged for each year between 2008 and 2019.

KEYWORDS: Downslope winds; Ensembles; Probability forecasts/models/distribution; Model evaluation/performance

1. Introduction

Downslope windstorms in the lee of mountains are high-impact weather events, where the peak winds have been known to exceed 50 m s^{-1} (Durran 1990). Direct wind damage from these events can be severe, and they can also drive wildfires, as occurred in the vicinity of Marshall, Colorado, on 30 December 2021 (Fovell et al. 2022). Therefore, in areas prone to these windstorms, including for example the Colorado Front Range and the Wasatch Front of Utah, accurate prediction with as much lead time as possible is important. Although mesoscale models are clearly capable of producing simulations representative of actual events (Colle and Mass 1998; Cao and Fovell 2016), previous studies regarding the predictability of downslope windstorms have led to differing, and sometimes contradictory, results.

The earliest attempt to predict downslope winds using a dynamical model driven by synoptic-scale numerical weather forecasts appears to be that of Klemp and Lilly (1975), who used a multilayer linear model to forecast windstorms in Boulder, Colorado. Anthes and Baumhefner (1984) cited this study as evidence that mesoscale phenomena generated by the interaction of synoptic-scale systems and known topographic barriers could potentially be forecast over much longer lead times than the predictability time scales estimated by Lorenz (1969) for similar size circulations in homogeneous isotropic turbulence. Echoes of the idea that terrain enhances mesoscale predictability continue to this day.

Updating the linear analysis in Klemp and Lilly (1975), Nance and Colman (2000) conducted two-dimensional nonlinear simulations to investigate the predictability of downslope windstorms at seven locations throughout the western United States. The performance of their model differed between cases with and without a mean-state critical layer (a layer in which the cross-mountain flow drops to zero or reverses direction relative to that at lower levels). When no mean-state critical level was present, the model seriously overpredicted the maximum gusts, often by more than 22 m s^{-1}. On the other hand, the errors in the maximum gusts were centered about zero in those cases with mean-state critical levels.

A mean-state critical level caps the upward propagation of energy carried by gravity waves launched by flow over a mountain barrier. Studies using semianalytic and numerical models (Smith 1985; Durran and Klemp 1987; Bacmeister and Pierrehumbert 1988) show that this capping effect generates a downslope wind response over a wide range of nondimensional critical level heights, suggesting therefore that downslope wind events occurring beneath mean-state critical levels may be comparatively easy to forecast. Indeed, Lawson and Horel (2015) found a 4-day predictability lead time in ensemble forecasts of the 1 December 2011 Wasatch windstorm, which occurred beneath a critical level near 350 hPa.

Another process that leads to strong downslope winds is wave breaking, where the flow above the mountain and lee slope becomes stagnant or reversed, forming a so-called "self-induced" critical level (Peltier and Clark 1979). Because there

Corresponding author: Dale R. Durran, drdee@uw.edu

DOI: 10.1175/WAF-D-22-0135.1

is a strong bifurcation of the downslope wind response depending on whether or not a mountain wave amplifies sufficiently to break, windstorms that occur under these conditions are particularly difficult to forecast (Doyle and Reynolds 2008). Both multimodel (Doyle et al. 2011) and initial-condition (Reinecke and Durran 2009) ensemble reforecasts of a wave-breaking event during the Terrain-Induced Rotor Experiment (T-REX) program further suggest that windstorms generated under such conditions can be quite sensitive to minor changes in the upstream flow and therefore have very limited predictability. More specifically, Fovell et al. (2022) found that 28-h lead time forecasts of the 30 December 2021 Marshall fire failed to predict the windstorm because a subtle error in the initial synoptic-scale analysis shifted an upstream region of sharp horizontal wind shear roughly 120 km too far to the north.

Strong downslope winds can also occur in the absence of a mean-state critical layer and without any wave breaking if a layer of strong static stability is present near mountain-top level in the cross-mountain flow, with weaker stability aloft. Although windstorms produced by such "static stability layering" appear to be more predictable than those generated by wave breaking, ensemble simulations of a static stability layering case from T-REX showed that the predictability lead time for that event was less than 12 h (Reinecke and Durran 2009).

These previous studies used a variety of approaches to study the predictability of downslope winds. Several investigations used 2D nonlinear models: Nance and Colman (2000) conducted single deterministic simulations of many different cases, whereas just a few cases were investigated by Doyle and Reynolds (2008) using initial-condition ensembles and by Doyle et al. (2011) using multimodel ensembles. Three-dimensional ensembles were employed in Reinecke and Durran (2009) and Lawson and Horel (2015), but as with the 2D ensemble studies, they focused on just a couple observed events.

One yet-to-be-employed methodology would be to conduct high-resolution 3D ensemble simulations of a very large number of real-world cases. Until recently it has been impractical to conduct such a study due to its high computational cost. However, archived data from high-resolution operational mesoscale ensembles have begun to appear. The National Center for Atmospheric Research (NCAR) mesoscale ensemble covered the contiguous United States at 3-km horizontal grid spacing and featured 10 ensemble members launched once per day over a 2.5-yr period from April 2015 to December 2017 (Schwartz et al. 2019).

In the following we use data from the NCAR ensemble to examine the predictability of downslope windstorms at 15 locations in the western United States in three contexts. We will compare the potential predictability of windstorms that develop in the presence and in the absence of a mean-state critical level by comparing their ensemble spread in section 3. We then investigate practical aspects of downslope windstorm prediction using two different approaches. The first, presented in section 4, evaluates yes/no windstorm hindcasts in terms of the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) using biased-adjusted wind speed output from the ensemble. Our second verification method, the continuous ranked probability score (CRPS), accounts for errors in the ensemble mean and penalizes for overly wide ensemble spread. The CRPS results are presented in section 5. We present our conclusions in section 6.

2. Methodology

a. NCAR ensemble description

The NCAR ensemble was a real-time convection-allowing ensemble designed and run by NCAR from 7 April 2015 to 30 December 2017. It was initialized at 0000 UTC each day and integrated to a model time of 48 h. The ensemble consisted of 10 WRF-ARW members run at 3-km horizontal grid spacing over the continental United States. All differences among the ensemble members were produced by varying the initial conditions; the dynamical core configuration and parameterization choices remained fixed among the ensemble members. Ensemble perturbations were generated from a continuously cycling ensemble Kalman filter (EnKF). Eighty (initially 50 until 2 May 2016) ensemble members with 15-km horizontal grid spacing were utilized in the data assimilation system. These members were updated every 6 h by assimilating a quality-controlled observational dataset. Every 24 h, at 0000 UTC, 10 of the members were downscaled to 3-km horizontal grid spacing and used to initialize the 48-h forecasts (Schwartz et al. 2019).

Output from all ensemble members is archived at NCAR, but not all fields are available. In particular, much archived output relates to convective indices useful for forecasting severe thunderstorms, and many of the three-dimensional prognostic variables have either been omitted or are saved at too coarse vertical resolution to adequately characterize the vertical structure of the cross-mountain flow. Therefore, we have necessarily restricted our use of this data to 10-m wind speed, which is output at a time resolution of 1 h.

The configuration of the ensemble is such that there are two ensemble forecasts for every forecast time within the verification period: one forecast from the ensemble initialized at 0000 UTC of the forecast day and one forecast from the ensemble initialized at 0000 UTC the day before. Therefore, for each forecast time, there are 20 verifying ensemble members: 10 from each of the two initialization times. The forecast lead times for the actual events vary throughout the day, from as short as 1 and 25 h in the case of an event at 0100 UTC to as long as 24 and 48 h for a forecast for 0000 UTC. (Technically, there is also an ensemble forecast with a lead time of 0 h, but since the analysis is not useful for forecasting purposes or for analyzing predictability, we have omitted it.)

One might envision that day-1 forecasts will show more predictability than forecasts for day 2 because of the shorter lead times. To investigate this, our analysis of the ensemble spread is conducted as a function of lead times from 1 to 48 h. Our CRPS analysis also treats the short and long lead time ensembles separately. On the other hand, our yes/no forecasts for the set of potential and realized downslope wind events are computed using all 20 members (10 from each initialization),

but are for multihour verification windows, thereby minimizing the dependence on the precise forecast lead time.

b. Station selection

Stations for the analysis were selected from a filtered Mesowest dataset (Horel et al. 2002). With the assistance of geographic information system (GIS) software and the U.S. Geological Survey's 3D Elevation Program (3DEP; Sugarbaker et al. 2014) terrain dataset, we selected 15 stations in the lee of mountains across the western United States, which experience high winds arriving from a downslope direction. All but two of these stations were selected from locations lying at the bottom of the sloping terrain in very close proximity to the mountain barrier, a configuration which includes well-known windstorm-susceptible sites near Boulder, Colorado, and Centerville, Utah. Some of the stations we identified are less well known, such as one in west Texas that experiences downslope flow from the Guadalupe Mountains and one in eastern Washington that experiences downslope flow from a ridge near the Hanford Site. In contrast to the other 13, the two southern California stations are positioned higher up the slope.

Our set of stations include CO109, which recorded one of the highest wind gusts during the December 2021 Marshall Fire windstorm, and SILSD, which is one of the windiest stations in San Diego County (Cao and Fovell 2016, 2018). A full list of stations, including the direction of the mountain-normal vector, is provided in Table 1. These stations are broadly distributed through the western United States, as shown in Fig. 1. The orientation of the stations relative to the local terrain and the mountain-normal vector is shown in Fig. 2.

All of these stations have observations of high sustained wind and gusts within the forecast period of the NCAR ensemble. The time series for each station was manually quality controlled to remove periods of obviously bad data. Little bad data were present, with the exception of a period at CO109 from 18 November 2015 to 13 March 2016. To determine a model forecast wind speed for comparison with these observations, we find the nearest two grid cells corresponding to the location of the observation station. Then, for each ensemble member, we take the value from the cell with the highest wind speed. Using the maximum over a pair of cells better accounts for potential mismatches between the true and model topography while still remaining local to the corresponding observation station.

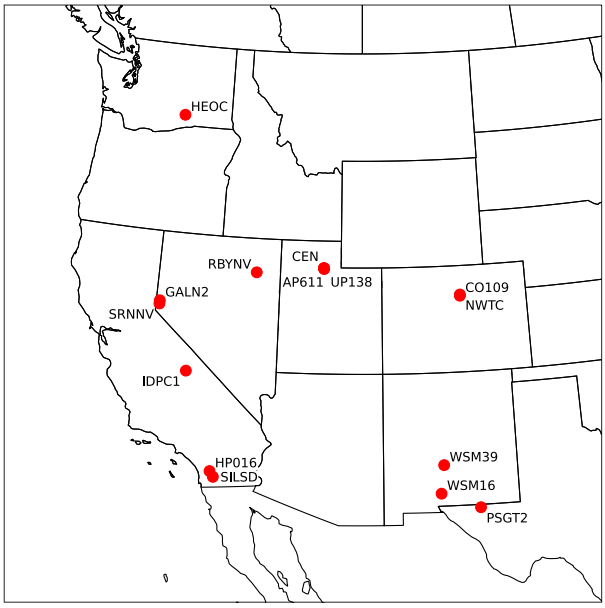

FIG. 1. Location of the 15 selected observation stations in the western United States.

c. Critical-level dataset

Because the NCAR mesoscale ensemble archive does not contain enough upper-level data to determine the presence and height of a critical level, this is estimated for each event using the ERA5 reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (Hersbach et al. 2020). Although the ERA5 data allow us to characterize the approximate

TABLE 1. List of the 15 selected stations by Mesowest ID and their general location. The direction of the mountain-normal vector and the mean forecast-minus-observation bias (see section 2d) at each station is also provided. Westside stations are italicized.

Mesowest station ID	Location	Mountain normal direction (°)	Mean bias (m s^{-1})
CO109	Boulder, CO	270	−3.6
NWTC	Boulder, CO	270	−1.9
SRNNV	Sierra Nevada Mountains, NV	230	−5.5
PSGT2	Guadalupe Mountains, TX	311	3.0
GALN2	Sierra Nevada Mountains, NV	277	4.8
IDPC1	Owens Valley, CA	268	3.8
WSM16	White Sands Missile Range, NM	294	−5.8
HEOC	Hanford Site, WA	213	−7.9
RBYNV	Ruby Mountains, NV	304	1.8
CEN	Centerville, UT	90	10.6
UP138	Centerville, UT	90	6.4
AP611	Centerville, UT	90	5.2
SILSD	San Diego County, CA	99	−5.0
HP016	San Diego County, CA	80	−5.1
WSM39	White Sands Missile Range, NM	61	−7.9

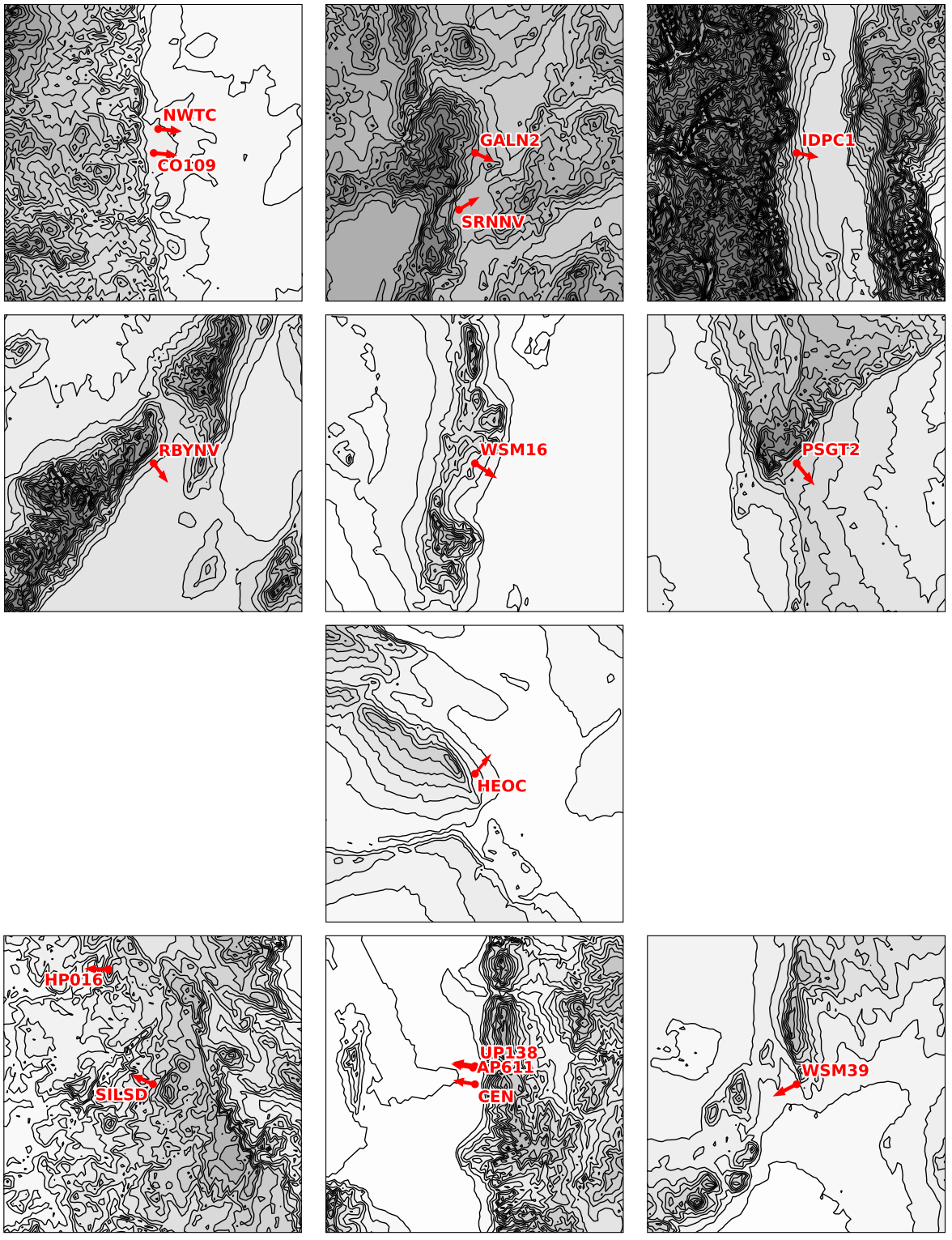

FIG. 2. Station location relative to terrain from the USGS 3DEP dataset. The top three rows show stations on the east side of mountain ranges, and the bottom row shows the westside stations. Filled contours indicate terrain height, with darker shades of gray indicating progressively higher terrain. Terrain contours are in 125-m intervals. The mountain-normal vector for each station is plotted as a red arrow. Each panel shows a 0.25° × 0.25° region.

elevation (or absence) of a critical level during each observed or forecast downslope windstorm, we have no information about the presence and elevation of any critical levels in the forecasts from each individual ensemble member.

One concern is whether or not the critical levels found in the ERA5 dataset are mean-state critical levels or are self-induced by breaking mountain waves. The ERA5 reanalysis was conducted using the ECMWF Integrated Forecast System at an effective horizontal grid spacing of 31 km (Hersbach et al. 2020), which is too coarse to resolve the velocity perturbations in breaking mountain waves. We, therefore, interpret all critical levels in the ERA5 data as mean-state critical levels.

d. Bias adjustment

Downslope windstorms tend to be very gusty (e.g., Durran 1990, Fig. 4.11), and the maximum gusts are more indicative of the potential storm impacts than the mean wind speed. The surface winds in the NCAR ensemble do not represent the maximum gusts carried to the surface by boundary layer turbulence, and they are also not forecasts for the precise location of each anemometer. To compare the modeled winds with the observed gusts, we evaluate a bias adjustment to the model forecasts.

To compute the bias at each station individually, we need to establish roughly compatible thresholds for the identification of downslope winds in both the forecasts and the observations. As a first preliminary step, we determine the average difference between the component-wise ensemble mean wind speeds:

$$\sqrt{\overline{u}_{\text{ens}}^2 + \overline{v}_{\text{ens}}^2}, \tag{1}$$

where $\overline{u}_{\text{ens}}$ and $\overline{v}_{\text{ens}}$ are the means of the west–east and south–north wind components, respectively, and the observed gusts over all observed downslope-directed winds at all stations during the NCAR ensemble period. For this comparison, we include all events with observed gusts greater than 10 m s^{-1} blowing from within ±45° of the vector normal to the main mountain ridge (the "windstorm quadrant"). On average, the model winds turn out to be 5.7 m s^{-1} less than the observed gusts. We then compute the bias individually at each station over the set of all events with winds within the windstorm quadrant having either observed gusts of at least 20 m s^{-1} or forecast wind speeds of at least 15 m s^{-1} in at least two ensemble members. Here the 15 m s^{-1} threshold uses the average overall observed minus forecast difference in downslope windiness, and is taken as a round-number approximation to (20 − 5.7) m s^{-1}. The two-ensemble-member threshold is chosen to be relatively nonrestrictive, thereby allowing a large sample size.

The bias could have been computed as the difference between the modeled winds and the observed wind gust in the direction of the mountain normal vector, as determined from the topographic data used by the model. But the model topography may differ from the actual topography, and even if they were identical, the mountain normal vector may not be perfectly aligned with the direction from which the strongest downslope winds typically blow at the anemometer. Therefore, if the wind direction is within the windstorm quadrant, we interpret the full wind speed as the forecast downslope wind. On the other hand, to account for cases with significant differences in wind direction, if the winds are blowing from outside the windstorm quadrant, we take the forecast speed as the ± magnitude of the projection of the model-simulated velocity onto the mountain normal vector.

To prevent nondownslope wind events from making it into our dataset, we apply another filter, which requires the cross-mountain wind component at or near the mountain top to be greater than 8 m s^{-1}. The magnitude of the criterion is motivated by Durran (1990), who notes "[Conditions] favorable for downslope winds occur when the wind is directed across the mountain (roughly within 30° of perpendicular to the ridgeline) and the wind speed at mountaintop level exceeds a terrain-dependent value of 7 to 15 m s^{-1}." This additional filter is implemented by taking a vertical cross section in the same plane as the mountain-normal vector. This cross section extends horizontally upstream 1° of latitude/longitude from the observation station. The maximum elevation within this cross section is taken as the mountaintop height. We then take the winds within the wedge extending horizontally ±30° of the mountain-normal vector and vertically between the mountaintop height and 2 km above the mountaintop height. If the maximum magnitude of these winds is greater than 8 m s^{-1}, the winds at this time are considered to be downslope winds and the data point is included in our final dataset; otherwise, it is neglected. Since above-surface fields are not available for the NCAR ensemble throughout most of the period, the cross sections are necessarily computed using ERA5 data as a proxy.

Applying this procedure to all forecast and/or observed events at each station, we obtain the average biases for the forecast winds minus the observations listed in Table 1. At four of the 15 stations, the magnitude of the bias is 6 m s^{-1} or greater. Two of these are overforecasts, with a maximum positive bias of 10.6 m s^{-1} at one of the stations near Centerville, Utah (CEN). The other two are underforecasts with a negative bias of −7.9 m s^{-1} at both the Hanford Site in Washington State (HEOC) and White Sands Missile Range, New Mexico (WSM39).

Examples of the evolution of the observed and bias-corrected forecast winds, together with the upper-level flow, are shown for a pair of events in Fig. 3. The velocity component aloft in the cross-mountain direction is plotted as a function of time and pressure level at CO109 (Fig. 3a) and SILSD (Fig. 3b), with the sign chosen so the downslope flow direction is positive. Values below −1 m s^{-1} indicate a critical level, which is shown as a black dashed line. Although theoretically stagnation (0 m s^{-1}) is sufficient for a critical level, we use 1 m s^{-1} of reversed flow to ensure a more robust critical-level signature. Note the absence of a critical level throughout the time series of CO109, while one is always present for the SILSD case.

The bias-adjusted winds at both stations are plotted as a function of time in the lower panels. Because the bias adjustment is negative, the weak winds prior to both events are adjusted upward by 3.6 and 5.0 m s^{-1} at CO109 and SILSD,

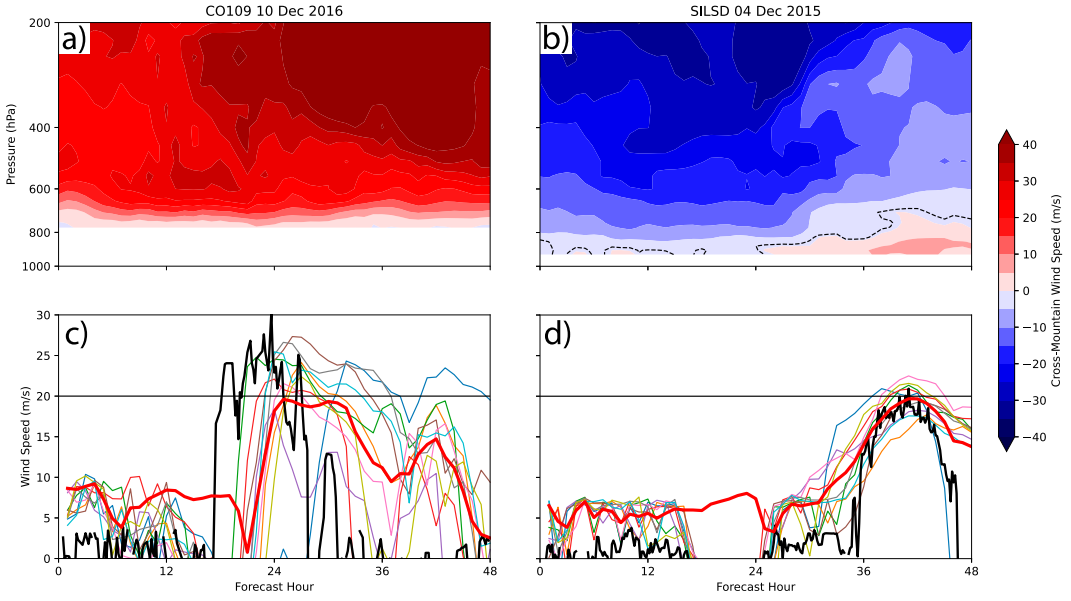

FIG. 3. Example windstorms at (a),(c) CO109 in Boulder, CO; and (b),(d) SILSD in Sill Hill, CA. (top) Time–height cross sections of the cross-mountain wind component (color fill) and critical level height (when present, dashed). (bottom) The bias-adjusted wind speeds for each ensemble member (thin colored curves), mean ensemble wind speed (thick red curve), and observed wind gusts (thick black curve).

respectively. The adjusted winds give a very good forecast of the high winds and their timing at SILSD (Fig. 3d), though the weaker wind period, which was not factored into the bias computation and is not used in our later downslope wind event scoring, is overforecast. At CO109 there are both timing and amplitude errors in the forecast (Fig. 3c), and once again an overestimate of the adjusted winds outside our period of interest. When forecasts are scored in section 4c, we will consider the influence of the size of the forecast window, i.e., the allowable timing error, on the verification statistics.

3. Ensemble spread in critical-layer and no-critical-layer events

a. *Computing ensemble spread*

If the WRF model and the data assimilation system were perfect, the potential limits to the predictability of these windstorms arising from the current level of observational uncertainty can be estimated by examining the spread among the ensemble members. Comparing the ensemble spread with the root-mean-square error (RMSE) of a set of forecasts is also one method for determining whether an ensemble is well calibrated (Fortin et al. 2014). Two slightly different formulas have been used to compute the average ensemble spread. The first takes the average over all forecasts (at a given lead time) of the standard deviations of the ensemble members for each forecast:

$$\mathrm{ES}_1 = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u_{j,i}-\overline{u}_j)^2}. \qquad (2)$$

The second takes the square root of the average variance of the ensemble members:

$$\mathrm{ES}_2 = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n-1}\sum_{i=1}^{n}(u_{j,i}-\overline{u}_j)^2}. \qquad (3)$$

In the preceding, n is the number of ensemble members, N is the number of verification times, $u_{j,i}$ is the wind speed of the ith ensemble member at verification time j, and $\overline{u}_j$ is the ensemble mean wind speed at verification time j. Note that the difference between the two metrics lies in when the square root operator is applied.

Only ES_2 is valid for determining whether or not an ensemble has the proper amount of spread (Murphy 1988; Fortin et al. 2014). However, ES_1 does have one advantage: it is the mean of some sample; namely, that of the standard deviations of the ensemble members, and therefore it is more natural to compute in conjunction with other summary statistics, such as the standard deviation and confidence intervals of the standard deviation distribution. Therefore, we use (2) to perform a statistical analysis of the spread to determine whether or not the difference in spread between no-critical-level and critical-level cases are significant, and we use (3) to assess the reliability of the ensemble by comparing it with the root-mean-square-error of the ensemble.

Our focus will be on the NCAR ensemble's performance on windstorm-relevant forecasts, specifically the union of the set of observed and the set of forecast events. For those cases satisfying our criterion on the cross-mountain wind component, we define the set of observed windstorms as those with gusts greater than 20 m s^{-1} within the windstorm quadrant,

and the set of forecast windstorms as those having at least two ensemble members with bias-corrected wind speeds greater than 20 m s^{-1} in the windstorm quadrant. The 20 m s^{-1} threshold was chosen as a round value that would yield lots of events while lying between the 15.6 m s^{-1} gust threshold for red-flag warnings in southern California (which also include relative humidity and fuel dryness criteria) and the 33 m s^{-1} gust threshold for NWS high wind warnings in many mountainous regions. Using thresholds defined by forecast values in individual ensemble members instead of an overall summary statistic like the ensemble mean is motivated by the possibility of a bifurcation in the strength of the simulated winds (Doyle and Reynolds 2008) in which the ensemble members exhibit either very strong winds or a minimal response. The total number of hours satisfying the preceding windstorm criteria across all stations is 13 154.

b. Comparing potential predictability

The windstorm-relevant forecasts are sorted into those with no critical level (no-CL), low-critical-level cases (low-CL) in which the critical level is lower than 8 km AGL, and high-critical level cases (high-CL) in which a critical level occurs between 8 km and 100 hPa. The distinction between the low-CL and high-CL categories is motivated by the findings in Nance and Colman (2000) that their 2D windstorm model performed better in critical-level cases in which the height of the critical level was lower than 8 km. They suggested that the low-CL case might be more predictable because the synoptic-scale flow would set the wave breaking level, whereas the processes determining the wave amplitude, including perhaps wave breaking, would be relatively independent of the elevation of the critical level in the high-CL cases. We consider critical levels above 100 hPa to be too high to have a significant impact on downslope wind development and, therefore, count these as no-CL cases.

The total number of cases in each category is plotted as a function of lead time in Fig. 4a. The dataset contains between 170 and 250 no-CL cases at every forecast lead time. Given the predominance of upper-level westerly flow in midlatitudes, it is not surprising that there are fewer low-CL cases, with 20–70 cases at each forecast lead time. The high-CL cases are rare, with around 20 cases per lead time. The bias-corrected ensemble-mean wind speeds are plotted as a function of forecast lead time in Fig. 4b. The no-CL ensemble mean has bias-corrected wind speeds that decrease slightly with time and average about 17 m s^{-1}. This ensemble average is less than the 20 m s^{-1} bias-corrected wind speed criterion for a simulated downslope wind because only two of the ten ensemble members need to exceed 20 m s^{-1} for the period to qualify as high-wind forecast. Both critical-level cases show a diurnal cycle with higher mean winds of 19–21 m s^{-1} dropping below 18 m s^{-1} around 0000 UTC, or equivalently, near forecast lead times 0, 24, and 48 h. No such diurnal signal is seen in the observations themselves (not shown). Lead times from a 0000 UTC forecast of 0, 24, and 48 h are the early evening hours in the western United States, so this spurious diurnal signal in the NCAR ensemble may be produced by errors in the boundary layer parameterization,

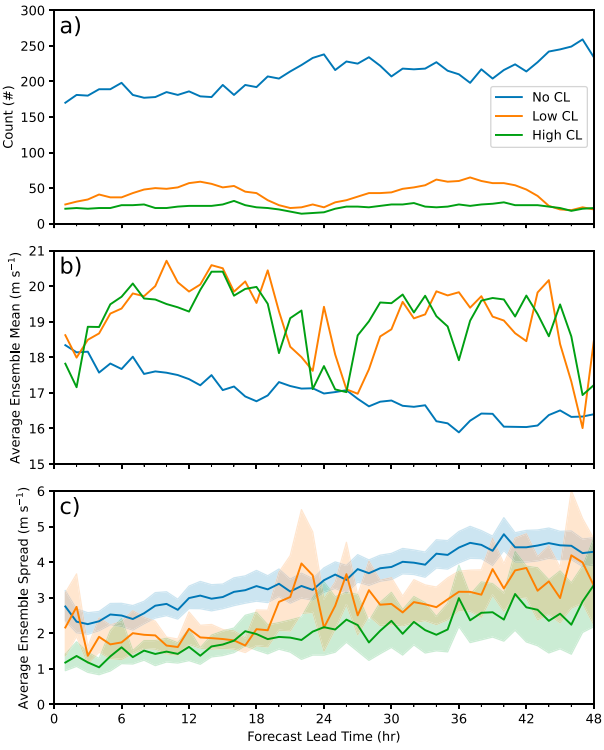

FIG. 4. Key parameters in no-CL events (blue), low-CL events (orange), and high-CL events (green) as a function of forecast lead time: (a) number of data samples, (b) ensemble-average mean bias-corrected wind speeds, and (c) ensemble-average standard deviation in forecast wind speed ES_1. Shading in (c) shows the 95% confidence intervals for the ensemble standard deviation.

although it is difficult to understand why such errors would not also impact the no-CL cases.

The ensemble spread for each category ES_1 is plotted as a function of forecast lead time in Fig. 4c. The spread for the no-CL cases increases almost monotonically with forecast lead time until about 1.5 days, and then remains nearly constant through the end of the forecast period. The ensemble spread for the high-CL cases is significantly smaller, just about half that for the no-CL cases. The ensemble spread for the low-CL cases generally lies between the other two cases, except that it increases sharply around 0000 UTC and exceeds that of the no-CL events around lead times of 2 and 24 h. Also plotted as shading in Fig. 4c are 95% confidence intervals for the statistical distributions of the ensemble spread at all lead times. The shading for the high-CL ensemble spread lies almost completely below that for the no-CL cases, suggesting the difference in the ensemble spread between the no-CL and high-CL cases is statistically significant. In contrast, at several times the confidence intervals for the ensemble spread in the low-CL cases overlaps, and even exceeds, that for the no-CL events, making the statistical significance of the differences in ensemble spread less clear cut between those windstorm conditions.

One important metric for an ensemble prediction system is whether it has the proper amount of spread to encompass the proper range of possible outcomes. One method of

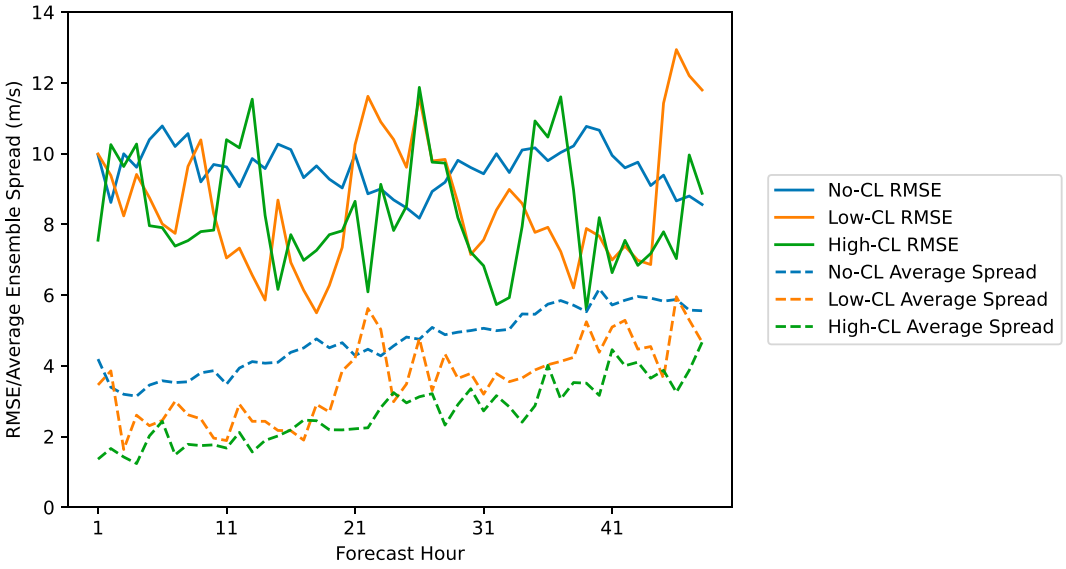

FIG. 5. Root-mean-square error (solid) and sample-size corrected ensemble spread $\sqrt{11/10}\,\mathrm{ES}_2$ (dashed), as a function of lead time for the no-CL (blue), low-CL (orange), and high-CL (green) cases.

determining whether an M-member ensemble has the proper amount of spread is check whether the sample-size corrected spread $[(M+1)/M]^{1/2}\mathrm{ES}_2$ matches the root-mean-square error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\overline{u}_j - \hat{u}_j)^2}, \tag{4}$$

where N is the number of verification times, $\overline{u}_j$ is the ensemble-mean bias corrected wind speed at verification time j, and $\hat{u}_j$ is the verifying observation. If the distribution of the ensemble members is unchanged when the observation is substituted for one of the ensemble members, the RMSE should match the sample-size corrected spread (Fortin et al. 2014).

The RMSE and sample-size corrected spread are plotted as a function of forecast lead time for the no-CL, low-CL, and high-CL cases in Fig. 5. The RMSE for all three types of events is much larger than the ensemble spread. Some lack of spread might be anticipated because initial condition ensembles are typically underdispersive (Buizza et al. 2005), and orographic precipitation forecasts generated using the NCAR mesoscale ensemble were shown to lack spread by Gowan et al. (2018). Nevertheless, with typical spread values less than one-quarter to one-half the RMSE, the ensemble is clearly failing to capture the true variability of downslope wind events. If the NCAR mesoscale ensemble were well calibrated, the larger ensemble spread in the no-CL cases would be a very strong indicator that the no-CL events have lower predictability than those cases with a critical level. But even in the current situation, in which the ensemble is underdispersive, the larger ensemble spread of the no-CL cases suggest that they are less predictable than those cases with critical levels,

Surprisingly, and in contrast to the spread, there is no systematic increase in RMSE with forecast lead time, although the low-CL RMSE does show a diurnal variation in phase with that superimposed on an upward trend with time for the low-CL average spread. Aside from this diurnal signal, the RMSE is a noisy function of lead time, ranging between 5 and 13 m s^{-1}. The mean RMSE over all lead times is 9.6, 8.6, and 8.3 m s^{-1} for the no-CL, low-CL, and high-CL cases, respectively.

4. Warn/no-warn forecasts

a. *Evaluation metrics*

We divided the roughly 2.5-yr period over which NCAR mesoscale ensemble forecasts are available into 6-h verification windows. Our results are not strongly sensitive to the size of the verification window, as will be discussed in section 4c. We evaluate forecast performance for each window period as a function of hits, false alarms, and missed events. All events are required to satisfy the criterion for cross-mountain winds. Downslope windstorms are deemed to have occurred when there was an observed gust of at least 20 m s^{-1} within the windstorm quadrant at any time during the forecast window. Individual ensemble members are deemed to have forecast a windstorm if they show bias-corrected surface winds in excess of 20 m s^{-1} within the windstorm quadrant at any time during the forecast window. Downslope wind warnings from the ensemble forecast system (actually hindcasts) are issued when at least N_T ensemble members forecast a windstorm, where N_T is the ensemble threshold number that ranges from 2 to 18 in our tests. (Recall that for this analysis the ensemble consists of 10 members from the forecast initialized at 0000 UTC on the date of the event and 10 from the forecast initialized at 0000 UTC the day before.)

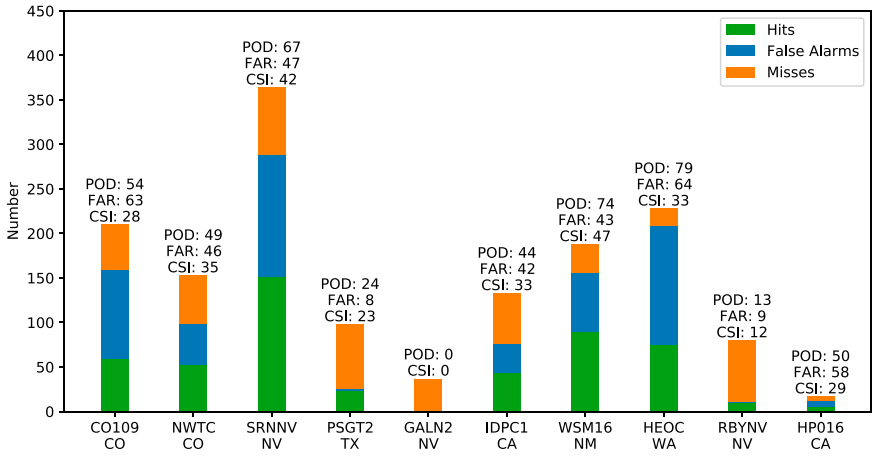

FIG. 6. Number of hits (green), false alarms (blue), and misses (orange) for no-CL cases at each station with more than 20 total hits, false alarms, and misses using $N_T = 4$ as the ensemble threshold number. Also noted are the corresponding values of POD, FAR, and CSI with the leading decimal point omitted.

Each window period is also categorized as a no-CL or critical-level (CL) event using winds from the ERA5 data at the closest grid point to the given station. We again define a critical level as occurring at the lowest elevation where the wind speed exceeds at least 1 m s^{-1} in the direction opposite the low-level cross-mountain flow. For some purposes, we will again distinguish between low- and high-critical-level cases, depending on whether the critical level is above or below 8 km AGL. If a critical level is present over 50% or more of the forecast window, we designate the case as a critical-level event.

The performance of our windstorm/no-windstorm forecast decision is evaluated using three metrics: the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI). These metrics have been utilized in many previous analyses of various binary (yes/no) forecasting problems (e.g., Schaefer 1990; Brooks and Correia 2018). Define a as the number of hits, where the ensemble forecasts a windstorm and one occurs (true positives), b as the number of false alarms when a windstorm is forecast but none is observed (false positives), and c as the number of missed events where the ensemble does not forecast a windstorm but one is observed (false negatives). The probability of detection is defined as the ratio of successfully forecast windstorms to all observed windstorms:

$$\mathrm{POD} = \frac{a}{a + c}. \tag{5}$$

The false alarm ratio is the ratio of false alarms to all forecast windstorms:

$$\mathrm{FAR} = \frac{b}{a + b}. \tag{6}$$

The critical success index (alternatively known as the threat score) is the ratio of successfully forecast windstorms to all cases where windstorms were a forecast issue:

$$\mathrm{CSI} = \frac{a}{a + b + c}. \tag{7}$$

The CSI can also be expressed as a combination of POD and FAR; it achieves a maximum value of unity when POD = 1 and FAR = 0, corresponding to a perfect forecast. CSI therefore provides one method of objectively combining POD and FAR into a single metric.

b. *Forecast performance*

The number of hits, false alarms, and misses for no-CL cases at each eastside station are charted in Fig. 6, using $N_T = 4$ as the ensemble threshold number. The POD, FAR and CSI computed from these values is noted above the bar for each station. Only stations with at least 20 total hits, false alarms, and misses are plotted. There are almost no downslope wind events for stations on the west side of the mountains unless a critical level is present, therefore the only westside station shown in Fig. 6 is HP016.

Forecast performance varies significantly between stations. The POD ranges between 0 and 0.79, while the FAR varies between 0.08 and 0.64. Reflecting the combined POD and FAR scores, the CSI ranges from very poor values of 0 to 0.47. Despite the modest $N_T = 4$ threshold, missed events are a huge problem at two Nevada stations GALN2 and RBYNV where the POD is 0 and 0.13, respectively, and mountain normal vectors are directed toward the southeast. Interestingly, SRNNV, which is located near GALN2 but has a mountain normal vector directed to the northeast, performs much better, with POD of 0.67 and the second highest CSI of the group. Overprediction is most serious in CO109 (Colorado) and HEOC (Washington), with FAR of 0.63 and 0.64, respectively. (The effect of increasing N_T on FAR will be discussed in connection with Fig. 9.)

Corresponding results for CL cases are shown are shown in Fig. 7. As noted in connection with Fig. 5a, there are fewer CL cases than no-CL cases. Yet in contrast to the no-CL

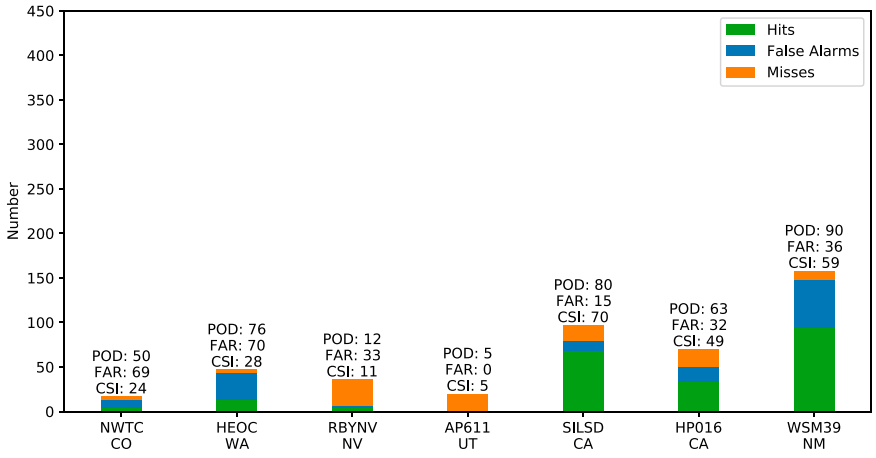

FIG. 7. As in Fig. 6, but for CL cases.

cases, which were essentially nonexistent for stations on the west side of mountain ranges, several eastside stations experienced some critical-level events. Synoptic charts for the eastside CL events often showed backing winds in connection with high-amplitude troughs that produced slight reversed flow at upper levels. Results are presented in Fig. 7 for all stations in which there were a total of at least 20 hits, false alarms, and misses during CL events.

The best results for the CL cases were obtained at SILSD (highest overall CSI of 0.70) and HP016 in the San Diego area, and at WSM39 in New Mexico (highest overall POD of 0.90). Despite the success of the 1 December 2011 Wasatch bora forecasts in Lawson and Horel (2015), the forecasts for the stations around Centerville, Utah, are very poor. Only one event was forecast at AP611 for any of the 20 observed events, yielding POD and CSI scores of 0.05. Similar results were obtained for stations CEN and UP138, although there were slightly less than 20 cases at these stations, so they are not plotted in Fig. 7. The preponderance of misses in the Wasatch events also contrasts with the comment in Lawson and Horel (2015) that "Further, subjective examination of other WRF high-resolution deterministic model forecasts for nascent downslope windstorms along the Wasatch Front suggests the model tends to overforecast their occurrence."

To examine any systematic differences in the performance of the no-CL and CL forecasts, all the cases in each category are aggregated together, and their average POD, FAR and CSI are shown in Fig. 8. When aggregated together, statistics for the low-CL and high-CL cases are almost identical, and therefore closely match the results for all CL cases combined. This similarity is robust across various choices of wind speed thresholds (not shown). The results for the aggregated no-CL cases are distinctly worse than the all-CL events with POD dropping from 0.61 to 0.52, FAR rising from 0.39 to 0.51, and CSI falling from 0.44 to 0.34. Analyzing the short (1–24 h) and long (25–48 h) forecast lead times separately, led to no significant differences in these metrics, which is surprising, but similar to the behavior of the RMSE shown in Fig. 5. In summary, the scores in Fig. 8 suggest the critical level cases are indeed easier to forecast than windstorms that occur in the absence of a critical level.

c. *Sensitivity to parameters*

1) NUMERICAL RESOLUTION

While it is not possible to change the grid spacing used in the archived WRF simulations, we can make a few observations about forecast performance at stations adjacent to wider or narrow ridges that are better or more poorly resolved in the WRF simulations. The axes in Fig. 2 span 0.25° in latitude and longitude, or equivalently about 6.5 grid points east–west and just over 9 grid points north–south. The station locations adjacent to the most poorly resolved topographic barriers are HEOC, WSM16, and WSM39, for which the CSI are 33, 47 (both no-CL), and 59 (CL), respectively. The performance at these stations is therefore approximately equal to, or better than, the average for all stations in their respective no-CL and CL categories (Fig. 8). On the other hand, the hindcasts for

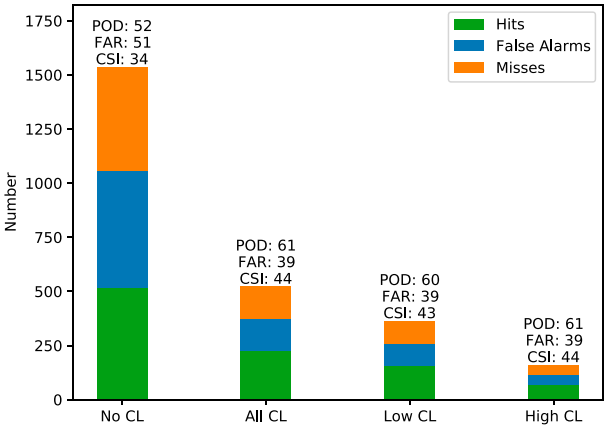

FIG. 8. Comparison of no-CL and CL events for all stations aggregated together. Statistics for CL events separated into low- and high-CL classes are also shown. The meaning of bar colors is equivalent to Fig. 6.

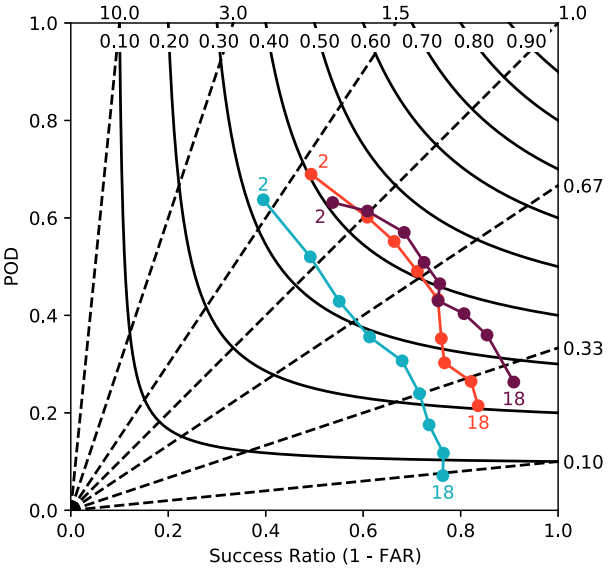

FIG. 9. Performance diagram for no-CL (light teal), low-CL (red), and high-CL (purple) cases as a function of ensemble member threshold. Endpoints corresponding to 2- and 18-member thresholds are indicated for each curve. Lines of equal CSI are solid black, with corresponding values along the inside top of the plot. Dashed black lines indicate bias values, with corresponding values denoted along the outside top and right of the plot.

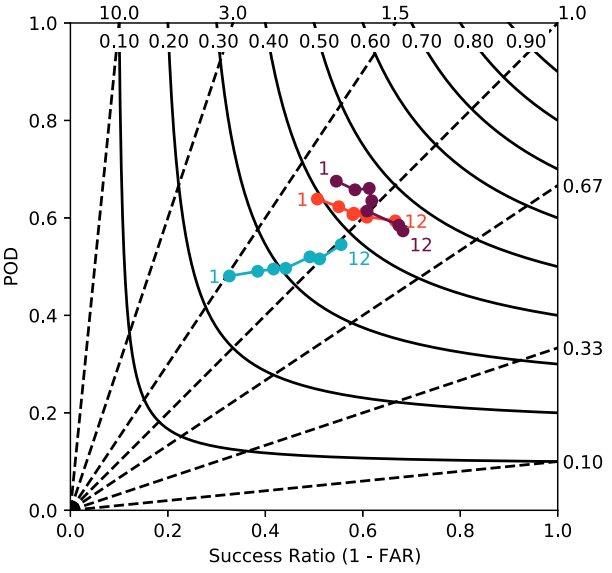

FIG. 10. As in Fig. 9, but values are plotted as a function of the size of the verification window for sizes of 1, 2, 3, 4, 6, 8, and 12 h: no-CL (light teal), low-CL (red), and high-CL (purple) cases. Endpoints corresponding to w_s of 1 and 12 h are indicated for each curve.

stations adjacent to some of the best resolved topography perform roughly equal to, or worse than, the relevant no-CL or CL full-station-set average. Such stations include NWTC, CO109, IDPC1 (all no-CL with CSI of 35, 28, and 33), and AP611 (CL with CSI of 5). Finer grid spacing should allow more accurate process studies of downslope winds at all the stations; in particular it might reduce the underforecast bias at the poorly resolved stations, which is -7.9 m s^{-1} at both HEOC and WSM39. But as suggested by this comparison, *increasing the numerical resolution is not guaranteed to actually improve bias-corrected forecasts.*

2) ENSEMBLE-MEMBER THRESHOLD, *NT*

The sensitivity of the results shown in Fig. 8 to the warning threshold N_T can be assessed using the performance diagram in Fig. 9, in which the POD is plotted for $N_T = 2, 4, \ldots, 18$ as a function of the success ratio, SR $= 1 -$ FAR (Roebber 2009). Reference curves consisting of isolines of CSI are plotted as solid curves in Fig. 9, along with dashed curves indicating the bias, defined as

$$\text{bias} = \frac{\text{number of forecast events}}{\text{number of observed events}} = \frac{a+b}{a+c} = \frac{\text{POD}}{\text{SR}}.$$

If a forecast system produced perfect results, the POD, SR, and bias would all be unity and would yield a point at the upper right corner of the performance diagram. For the no-CL forecasts (light teal), choosing $N_T = 4$ gives both the best CSI and the least bias. Similarly, $N_T = 4$ is optimal for the low-CL cases (red), and it is also a good choice for the high-CL cases (purple)—although switching to $N_T = 6$ would give a slightly better CSI, $N_T = 4$ gives the lowest bias. It is interesting to note that the best N_T threshold requires a windstorm in only 1/5 of the total ensemble members. *Waiting for a majority of the ensemble members to indicate a windstorm before issuing a warning would not optimize the CSI.* Fig. 9 also shows the same key takeaways evident in Fig. 8 continue to hold over the range $2 \le N_T \le 10$; namely, that no-CL cases are harder to forecast than CL cases, and that low- and high-CL forecasts exhibit similar error.

3) WIDTH OF VERIFICATION WINDOW

One might expect the FAR to drop (and SR to increase) as the size of the verification window w_s increases, and as shown in Fig. 10, this is indeed the case, with the most significant improvements occurring before w_s rises to 6 h. Interestingly, as w_s increases from 1 to 12 h the POD improves for no-CL cases, but it degrades for both low-CL and high-CL events. The drop in POD with increasing w_s suggests that unforecast events are appearing at the beginning or end of the verification window, and it is surprising that this might be more of a problem in CL cases, which are nominally expected to be less sensitive than no-CL cases to small changes in the synoptic-scale flow. The CSI for the no-CL cases increases substantially with larger w_s because of the combined improvements in both the POD and SR. In contrast, the improvement in SR is largely canceled by the degradation in POD in both CL cases, thereby leaving the CSI almost independent of w_s. The bias is reduced for all three categories as w_s increases from 1 to 6 h, and is almost perfect at 6 h. There is little further change in the bias in the no-CL category as w_s is further increased to 8 and 12 h, whereas the bias shifts toward under forecasting events in

both CL categories. This shift toward underforecasting with increasing w_s in CL cases would again be consistent with the appearance of unforecast events at the beginning or end of the verification window.

d. Comparison with general high-wind forecasts

As discussed in the introduction, Anthes and Baumhefner (1984) suggested that terrain-induced mesoscale circulations might be predictable at longer lead times than many other mesoscale phenomena because the synoptic-scale flow can be forecast comparatively far in advance and the terrain itself can be specified with extremely high accuracy. Moreover, they cited an early attempt at downslope wind prediction (Klemp and Lilly 1975) as key evidence for this possibility. It is beyond the scope of this study to conduct a rigorous comparison between the predictability of downslope winds and high winds in non-mountainous regions, by for example, selecting another 15 stations in flat terrain and repeating the above analysis. Nevertheless, one perspective from which we can view this 40-yr-old hypothesis is provided by the performance diagram in Fig. 11, which again shows forecast scores for our optimal $N_T = 4$, $w_s = 6$-h ensemble predictions for the no-CL, low-CL, and high-CL categories, plotted along with the scores averaged over all National Weather Service (NWS) high-wind warnings and over each fiscal year from 2008 to 2019.

The vast majority of the NWS high wind cases do not involve downslope winds and may be taken as an indication of the operational skill in forecasting high-wind events that are not downslope. Although the NWS CSI scores are tightly clustered between about 0.34 and 0.41, these forecasts improved significantly over this period in that the average forecast lead time gradually increased from 6 to 12 h (Durran 2020). The NWS speed threshold criteria for high-wind warnings in nonmountainous regions are typically sustained winds greater than 40 mph (18 m s^{-1}) or gusts greater than 58 mph (26 m s^{-1}); in most mountainous regions the thresholds for winds and gusts increases to 50 and 75 mph (22 and 33 m s^{-1}), respectively. The thresholds for the NWS high wind events are therefore higher than our bias adjusted 20 m s^{-1} gust threshold—their events are more extreme.

As apparent in Fig. 11 the annual averaged CSI for the NWS high-wind forecasts was equal to or better than the 0.34 value obtained with our ensemble for the no-CL events, but slightly worse than the 0.43–0.44 values achieved in the CL cases. In contrast to the ensemble-based warnings, which are unbiased, the NWS forecasts are biased toward overforecasting the number of events. On balance, this rough comparison suggests that downslope winds developing in the absence of a critical layer are harder to forecast than an "average" extreme high-wind event. The CL cases, on the other hand, may be marginally easier to forecast.

5. Evaluating the entire ensemble distribution

The preceding evaluates the success with which downslope windstorm warnings can be issued based on various thresholds in the number of ensemble members exceeding a set value for the bias-adjusted wind speed. While very relevant to the operational forecast process, this approach does not fully assess the predictive information in the full ensemble. As a complimentary metric, we evaluate the continuous ranked probability score CRPS (Hersbach 2000). The CRPS penalizes both narrow distributions with an incorrect mean (i.e., an overly confident, but incorrect, forecast) and overly broad distributions, even if they have the correct mean (i.e., extremely uncertain forecasts).

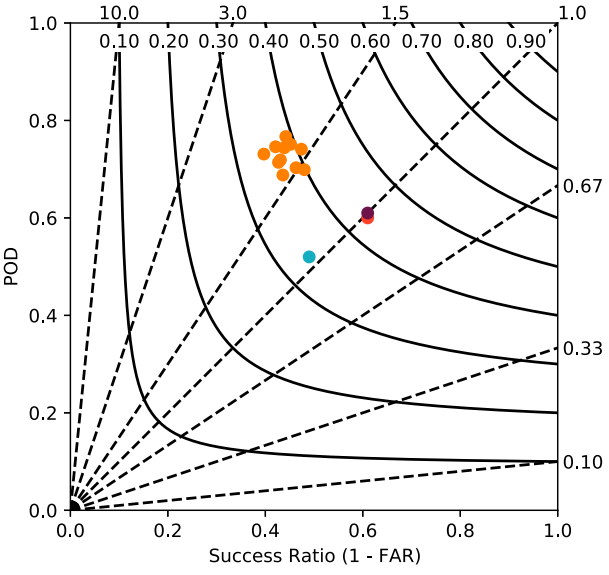

FIG. 11. Performance diagram, as in Figs. 9 and 10, but for National Weather Service high-wind warnings for the period from October 2007 to September 2019. Values for each fiscal year (1 Oct–30 Sep) are plotted in orange. Corresponding values for no-CL, low-CL, and high-CL cases are plotted in teal, red, and purple, respectively.

For a forecast variable x, the CRPS measures how well the cumulative distribution function (CDF) of the ensemble $P(x)$ matches the CDF of the observation $P_a(x)$, and is computed as

$$\text{CRPS} = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 \, dx. \tag{8}$$

Denoting the observed value of x as x_a, the probability density function (PDF) of the observation is $\rho_o(x) = \delta(x - x_a)$, and the CDF of the observation is

$$P_a(x) = \int_{-\infty}^{x} \rho_o(u) \, du = H(x - x_a), \tag{9}$$

where $H(x)$ is the Heaviside function:

$$H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases} \tag{10}$$

The PDF of an n-member ensemble forecast is given by

$$\rho_e(x) = \sum_{i=1}^{n} \frac{1}{n} \delta(x - x_i), \tag{11}$$

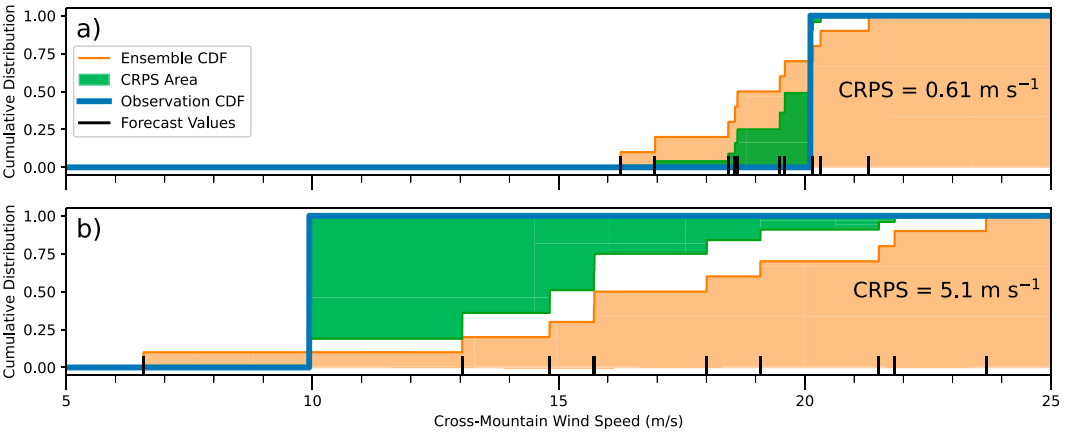

FIG. 12. (a),(b) Schematic of the calculation of CRPS for the example 10-member ensembles in two windstorm cases. Forecast wind speeds x for each ensemble member are indicated by the upward black ticks along the x axis. Note two ensemble members forecast wind speeds of 15.7 m s^{-1} in (b). The CDFs for the ensemble forecasts $P(x)$ are plotted in orange with shading; the CDF for the observations $P_a(x)$ is plotted as the heavy blue line. The green area illustrates the integral of $[P(x) - P_a(x)]^2$. The CRPS values are noted in each plot.

where x_i is the value forecast by the ith ensemble member. Therefore,

$$P(x) = \int_{-\infty}^{x} \rho_e(u)\, du = \sum_{i=1}^{n} \frac{1}{n} H(x - x_i). \tag{12}$$

Note the CDFs are dimensionless, so the CRPS inherits the dimensions of the forecast variable x owing to the integral in (8). Lower CRPS scores are better.

To gain intuition about the numerical values of the CRPS in our application, the CDFs and the integrand in (8) are illustrated for two windstorm cases in Fig. 12. Here, as in the ensemble spread analysis, the short and long lead times are treated as two separate 10-member ensembles with ensemble-forecast windstorms defined as 2 (out of 10) members greater than 20 m s^{-1} within the windstorm quadrant. One case, with a CRPS of 0.62 m s^{-1}, has narrow spread and only a small error in the ensemble mean. The other case, with a CRPS of 5.1 m s^{-1}, has a larger spread and much larger error in the ensemble mean—although the observation does lie within the ensemble spread.

The frequency distribution of CRPS scores for the no-CL, low-CL, and high-CL categories are shown in Fig. 13. These values are normalized by the total number of events in each category. There are no CRPS values below 0.1 m s^{-1}, but the number of good forecasts with CRPS $\leq$ 0.5 ramps up very rapidly in all three cases.

Once again, the performance on the no-CL cases is the worst, with a median CRPS of 4.14 m s^{-1} compared to the low-CL events with median CRPS of 3.42 m s^{-1}, and the high-CL events with a median of 3.90 m s^{-1}. Similarly, the third-quartile value is lowest for the low-CL category and highest for the no-CL category. Finally, the mean CRPS scores give the same relative ranking for each category: no-CL 5.95, low-CL 5.15, and high-CL 5.64 m s^{-1}. Analyzing CRPS as a function of forecast lead time led to no significant results (not shown).

The CRPS measures the quality of the full ensemble, and it is interesting that the slightly worse performance of the high-CL ensemble forecasts relative to the low-CL cases is consistent with the CSI scores obtained using ensemble number thresholds above $N_T = 4$ (Fig. 9).

6. Discussion and conclusions

The relatively long predictability lead times for synoptic-scale weather patterns have long been thought to enhance the predictability of terrain-induced mesoscale flows. Downslope windstorms have been suggested as a prime example of high-impact mesoscale weather events that inherit enhanced predictability from the synoptic scale (Anthes 1984). On the other hand, more recent research has suggested downslope

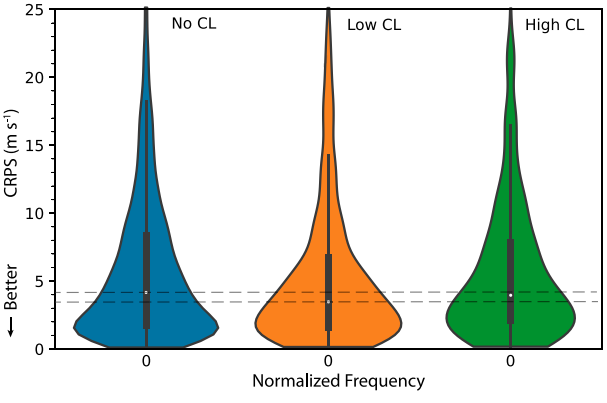

FIG. 13. Violin plots of the CRPS distributions for the no-CL (blue), low-CL (orange), and high-CL (green) cases, normalized by the total number of events in each case. White dots indicate distribution medians, and vertical black bars span the distance between the first and third quartiles. The black whiskers represent points that fall within 1.5 times the interquartile range. Thin dashed reference lines pass through the no-CL and low-CL medians.

windstorms have low predictability because they are highly sensitive to small changes in the large-scale flow (Reinecke and Durran 2009). Nevertheless, the 1 December 2011 downslope wind event along the Wasatch range provides an example of a downslope windstorm with excellent long-lead-time predictability (Lawson and Horel 2015). A mean-state critical level was present during the Wasatch event, and previous theoretical and modeling work suggests there may be higher predictability when such a critical level is present. In this study we have tested the hypothesis that the presence of a mean-state critical level tends to enhance the predictability of downslope windstorms, while also investigating the predictability of downslope windstorms in general.

The key and unique tool used in our analysis was the archive of NCAR mesoscale-ensemble-system forecasts, which provided 10 members integrated forward for 48 h from different initial conditions once daily on a 3-km horizontal grid over a period of about 2.5 years. Bias-corrected surface wind forecasts from this archive were compared against observations at 15 downslope-windstorm-prone observational stations across the western United States. Using a 6-h verification window, our dataset included over 2000 observed and/or forecast windstorm cases, giving a large sample size and allowing us to account for overprediction of nonevents by the model. We characterized the predictability in three ways: by evaluating the ensemble spread, the performance of yes/no windstorm forecasts, and the continuous ranked probability score (CRPS).

Comparing the ensemble spread allows us to estimate the potential predictability of events with and without mean-state critical levels in a hypothetical situation with a perfect forecast model and data assimilation system. The spread in the no-CL cases was roughly double that in the high-CL cases (critical level above 8 km AGL), with the spread for the low-CL category lying in between. Although there were relatively few high-CL cases, the difference in spread between them and the no-CL cases was significant. The significance of the difference in spreads between the low-CL and the no-CL cases was less clear cut. Nevertheless, the overall results suggest greater potential predictability for downslope winds that occur beneath a mean-state critical than for those occurring when no critical level is present. In all categories (no-CL, low-CL and high-CL), the ensemble spread was much smaller than the RMSE of the forecasts, indicating that the ensembles are underdispersive, a common situation for ensembles generated solely by perturbing the initial conditions.

The CRPS, which provides a measure of the accuracy of the full ensemble, penalizes both for errors in the mean and for overly wide spread. Our comparison of the CRPS scores found that ensemble forecasts for no-CL events are worse than those for cases with a critical level. But in contrast to the situation with the potential predictability estimated from ensemble spread, the best CRPS scores were for the low-CL events. The average CRPS scores for high-CL events were worse than those for low-CL events, but still better than for the no-CL cases.

We tested criteria for issuing a windstorm warning based on the number of ensemble members showing bias-corrected winds greater than 20 m s^{-1}. The best performance was obtained by setting a threshold in which just 4 of 20 ensemble members were required to trigger a warning. Dividing the roughly 2.5-yr period of the NCAR mesoscale ensemble archive into 6-h verification windows, our dataset included roughly 2000 windstorm forecasting challenges distributed across the 15 observations sites. The probability of detection (POD), false-alarm ratio (FAR), and critical success (CSI) indices for these hindcasts were all worse for the no-CL than for the CL cases. In particular, the average CSI for the roughly 1500 no-CL cases was 0.34, whereas it was 0.44 for the approximately 500 CL cases. For comparison, the CSI for 1-yr averages of all NWS high-wind warnings throughout the United States from 2008 to 2019 ranged from 0.34 to 0.41.

One interesting avenue for further study would be to test whether our results are too pessimistic for applications involving red-flag warnings in wildfire meteorology. Our best and third best results were obtained for the only two stations that are located on lee slopes rather than flat terrain at the base of the topography. It is possible that downslope winds may be easier to forecast at locations farther up the lee slope than at the foot of the mountain—particularly if the goal is to predict whether high winds will occur at any *unspecified* location along that slope.

The factors that determine whether downslope winds penetrate to the base of the lee slope involve difficult to simulate boundary layer dynamics. The need to remove preexisting cold-air pools is one factor. Mayr and Armi (2008) note that for deep Alpine south foehn, the virtual potential temperature of the air crossing the ridge needs to exceed that preexisting in the Inn Valley. An accurate representation of momentum loss in the boundary layer is also important. Higher values of drag and surface roughness reduce not only the surface wind maxima, but also the distance the high wind region extends down the lee slope (Richard et al. 1989; Miller and Durran 1991; Cao and Fovell 2018).

An example where sustained downslope winds do not penetrate all the way down the lee slope to the valley floor in the vicinity of Geyserville and Healdsburg, California, is shown in Fig. 14a. On 27 October 2019, when the Kincade fire was spreading in the area, the time series of wind observations at Healdsburg Hills (Fig. 14b), located at the blue dot in Fig. 14a, show north-northeast sustained winds maximizing at 30 m s^{-1} and gusts exceeding 40 m s^{-1}. In contrast, the winds nearby at Red Fan (Fig. 14c), close to the base of the ridge at the red dot in Fig. 14a, are much weaker, with sustained winds maximizing at 8 m s^{-1} and gusts to 15 m s^{-1}. There is no cold pool isolating the valley station from stronger winds aloft; at 1200 UTC both Red Fan and Healdsburg Hills lie on the same 289-K dry adiabat. This must, therefore, be a case where the windstorm dynamics (including surface friction) do not bring strong winds all the way down the lee slope.

Finally, we note that the two southern California stations located up on the lee slopes, for which forecasts performed quite well, SILSD (CSI 0.70) and HP016 (CSI 0.49), contributed significantly to the overall superiority of the critical-level-event forecasts. It is possible, therefore, that some of the improved predictability we found for cases occurring below a

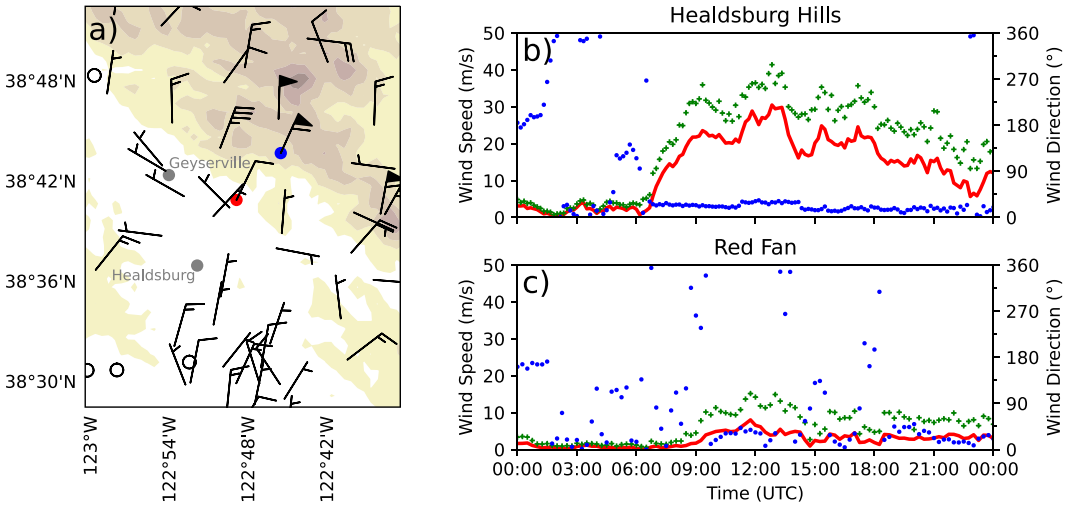

FIG. 14. Kincade fire winds from 27 Oct 2019. (a) Sustained winds at 1300 UTC [full barbs = 10 kt (1 kt ≈ 0.51 m s^{-1})], terrain contours every 250 m. Time series of sustained winds (red), gusts (green), and direction (blue) at (b) Healdsburg Hills (blue dot) and (c) Red Fan (red dot).

mean-state critical level, relative to cases with no critical level, actually arose from our particular choice for the station locations, with all no-CL stations down at the base of the mountain.

Acknowledgments. Comments from Paul Schlatter and two anonymous reviewers allowed us to significantly improve the manuscript. We benefited from discussions and data from Alexander Tardy and Brandt Maxwell. This research was supported by National Science Foundation Grant AGS-1929466.

Data availability statement. All NCAR ensemble data used in this study are publicly available from the NCAR Research Data Archive at https://doi.org/10.5065/D68C9TZ3. Mesowest data used in this study are provided by Synoptic Data's Mesonet API at https://synopticdata.com.

REFERENCES

Anthes, R. A., 1984: Predictability of mesoscale meteorological phenomena. *Predictability of Fluid Motions*, G. Holloway and B. J. West, Eds., American Institute of Physics, 247–270.

——, and D. P. Baumhefner, 1984: A diagram depicting forecast skill and predictability. *Bull. Amer. Meteor. Soc.*, **65**, 701–703, https://doi.org/10.1175/1520-0477-65.7.701.

Bacmeister, J. T., and R. T. Pierrehumbert, 1988: On high-drag states of nonlinear stratified flow over an obstacle. *J. Atmos. Sci.*, **45**, 63–80, https://doi.org/10.1175/1520-0469(1988)045<0063:OHDSON>2.0.CO;2.

Brooks, H. E., and J. Correia, 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Cao, Y., and R. G. Fovell, 2016: Downslope windstorms of San Diego County. Part I: A case study. *Mon. Wea. Rev.*, **144**, 529–552, https://doi.org/10.1175/MWR-D-15-0147.1.

——, and ——, 2018: Downslope windstorms of San Diego County. Part II: Physics ensemble analyses and gust forecasting. *Wea. Forecasting*, **33**, 539–559, https://doi.org/10.1175/WAF-D-17-0177.1.

Colle, B. A., and C. F. Mass, 1998: Windstorms along the western side of the Washington Cascade Mountains. Part II: Characteristics of past events and three-dimensional idealized simulations. *Mon. Wea. Rev.*, **126**, 53–71, https://doi.org/10.1175/1520-0493(1998)126<0053:WATWSO>2.0.CO;2.

Doyle, J. D., and C. A. Reynolds, 2008: Implications of regime transitions for mountain-wave-breaking predictability. *Mon. Wea. Rev.*, **136**, 5211–5223, https://doi.org/10.1175/2008MWR2554.1.

——, and Coauthors, 2011: An intercomparison of T-REX mountain-wave simulations and implications for mesoscale predictability. *Mon. Wea. Rev.*, **139**, 2811–2831, https://doi.org/10.1175/MWR-D-10-05042.1.

Durran, D. R., 1990: Mountain waves and downslope winds. *Atmospheric Processes over Complex Terrain*, *Meteor. Monogr.*, No. 23, Amer. Meteor. Soc., 59–81.

——, 2020: Can the issuance of hazardous-weather warnings inform the attribution of extreme events to climate change? *Bull. Amer. Meteor. Soc.*, **101**, E1452–E1463, https://doi.org/10.1175/BAMS-D-20-0026.1.

——, and J. B. Klemp, 1987: Another look at downslope winds. Part II: Nonlinear amplification beneath wave-overturning layers. *J. Atmos. Sci.*, **44**, 3402–3412, https://doi.org/10.1175/1520-0469(1987)044<3402:ALADWP>2.0.CO;2.

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeor.*, **15**, 1708–1713, https://doi.org/10.1175/JHM-D-14-0008.1.

Fovell, R. G., M. J. Brewer, and R. J. Garmong, 2022: The December 2021 Marshall fire: Predictability and gust forecasts from operational models. *Atmosphere*, **13**, 765, https://doi.org/10.3390/atmos13050765.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

——, and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Horel, J., and Coauthors, 2002: MesoWest: Cooperative Mesonets in the western United States. *Bull. Amer. Meteor. Soc.*, **83**, 211–226, https://doi.org/10.1175/1520-0477(2002)083<0211:MCMITW>2.3.CO;2.

Klemp, J. B., and D. R. Lilly, 1975: The dynamics of wave-induced downslope winds. *J. Atmos. Sci.*, **32**, 320–339, https://doi.org/10.1175/1520-0469(1975)032<0320:TDOWID>2.0.CO;2.

Lawson, J., and J. Horel, 2015: Ensemble forecast uncertainty of the 1 December 2011 Wasatch windstorm. *Wea. Forecasting*, **30**, 1749–1761, https://doi.org/10.1175/WAF-D-15-0034.1.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21A**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Mayr, G. J., and L. Armi, 2008: Föhn as a response to changing upstream and downstream air masses. *Quart. J. Roy. Meteor. Soc.*, **134**, 1357–1369, https://doi.org/10.1002/qj.295.

Miller, P. P., and D. R. Durran, 1991: On the sensitivity of downslope windstorms to the asymmetry of the mountain profile. *J. Atmos. Sci.*, **48**, 1457–1473, https://doi.org/10.1175/1520-0469(1991)048<1457:OTSODW>2.0.CO;2.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **114**, 463–493, https://doi.org/10.1002/qj.49711448010.

Nance, L. B., and B. R. Colman, 2000: Evaluating the use of a nonlinear two-dimensional model in downslope windstorm forecasts. *Wea. Forecasting*, **15**, 715–729, https://doi.org/10.1175/1520-0434(2000)015<0715:ETUOAN>2.0.CO;2.

Peltier, W. R., and T. L. Clark, 1979: The evolution and stability of finite-amplitude mountain waves. Part II: Surface wave drag and severe downslope windstorms. *J. Atmos. Sci.*, **36**, 1498–1529, https://doi.org/10.1175/1520-0469(1979)036<1498:TEASOF>2.0.CO;2.

Reinecke, P. A., and D. R. Durran, 2009: Initial-condition sensitivities and the predictability of downslope winds. *J. Atmos. Sci.*, **66**, 3401–3418, https://doi.org/10.1175/2009JAS3023.1.

Richard, E., P. Mascart, and E. C. Nickerson, 1989: The role of surface friction in downslope windstorms. *J. Appl. Meteor.*, **28**, 241–251, https://doi.org/10.1175/1520-0450(1989)028<0241:TROSFI>2.0.CO;2.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, https://doi.org/10.1175/BAMS-D-17-0297.1.

Smith, R. B., 1985: On severe downslope winds. *J. Atmos. Sci.*, **42**, 2597–2603, https://doi.org/10.1175/1520-0469(1985)042<2597:OSDW>2.0.CO;2.

Sugarbaker, L. J., E. W. Constance, H. K. Heidemann, A. L. Jason, V. Lukas, D. L. Saghy, and J. M. Stoker, 2014: The 3D elevation program initiative—A call for action. USGS Circular 1399, 35 pp., https://pubs.usgs.gov/circ/1399/pdf/circ1399.pdf.