General feedback

Overall, this manuscript is well written, well-structured and appears carefully worked through with nice looking figures. The work describes a new 4km, 20y mesoscale dataset covering North America with extensive validation using met towers and surface stations and ensemble analysis for a selected period.

Downscaling of global reanalysis models using mesoscale models like WRF is well covered in the literature as well as the improvements it provides relative to the global models. Hence, the novelty of the approach in this manuscript may be disputed given that it has poorer resolution compared to the previous work of e.g. Draxl et al. (2015). However, the open access to the large dataset and the extensive validation effort including ensemble analysis justifies the publication.

General comments

In general, I would like to question if the selected validation metrics for wind speed (r, RMSE, rRMSE, OVL) provide sufficient complementary insight. In my view, these metrics overlap too much in what they measure and none of them allow for distinction between systematic errors (biases) and fluctuating errors. I suggest including a simple metric like mean (bias) error to cover this important aspect and re-reconsider if each of the other metrics contribute enough additional insight to remain in the paper. A metric should be included only if characteristic error structures can be inferred from it – to move beyond being merely descriptive.

I suggest reducing the mostly summarising parts (section 3) with long descriptions and lists of numbers in the text. Please also consider additional summary table(s) for better overview and readability.

The paper should include consideration/discussion of the effect of not accounting microscale effects. A 4km model effectively resolves scales from 20-30km and up. How is this expected to affect presented results, when validating the model against measurements that include significant effects on finer scales, which may be very strong at 10m agl.?

Argumentation that the selected ensemble runs represent model uncertainty should be strengthened, this currently is an implied assumption. Does the spread across the selected and boot-strapped ensamples really represent actual model uncertainty?

The limitations and uncertainty of the observations used in the validation should be discussed either in section 2.2 or section 4.

Some detailed comments

Page 2, line 63: It should be mentioned here that ERA5 is initial/boundary model in addition to the info in table 1, on page 5.

Page 10, line 223: Explain “internal variability” and “structure uncertainty” in more detail, and why 10 and 6 ensemble members , respectively, was decided upon.

Page 14-15, fig. 3: A legend is missing for the plots.

Page 15, line 340: Interpolation in wind direction simply requires conversion of wind direction to components which may be interpolated similar to the wind speeds, and then converted back to wind directions.

Page 25, line 489: Friction velocity is denoted using u_* and not u*.

Page 26, line 503: Explain why “high friction velocities correspond to weaker winds”