Review for EGUsphere: egusphere-2022-355
"Evaluation of a cloudy cold-air pool in the Columbia River Basin in different versions of the HRRR model" - *by Bianca Adler et al.*

General Comments

This manuscript evaluates the development, evolution, and dissipation of a cold-air pool (CAP) in different versions of the HRRR model. Specifically, they consider HRRR v1 (CTL), a modified version of CTL that used model improvements from the WFIP2 field campaign (EXP), and HRRR v4 (v4fp1) that included several modifications including to the MYNN PBL scheme. Each of these was applied to two different grid spacings. The model data were compared against in situ and remote observations in and around the Columbia River Basin. The paper is generally well-written and the topic is relevant to EGU readers. The authors motivate the need for improved model representation of, e.g., winds in the considered conditions due to the reduced efficiencies in wind turbines. I have a few issues with the manuscript as presented.

First, I wonder why the authors use a cold-start approach starting at exactly 00 UTC with a 3-hour spin-up? The current method results in 3-hour periods of empty data in, e.g., Fig. 3 (which the authors note). Is this a limitation of the RAP output used to supply the initial conditions? If not, did the authors consider an approach where they initialize the model at 21 UTC on the day prior, use the same 3-hour spin-up, then run for 24-hours from 00–00 UTC? Even if RAP necessitated the 00 UTC initialization, why not initialize at the same 00 UTC and run for 24 hours from 03–03 UTC and avoid such gaps? Or was this all because the CTL and EXP were run this way in the past and the authors did not want to duplicate efforts? It seems like an unnecessary constraint. Second, and related, did the authors consider running (at least the v4fp1) a simulation for an entire CAP duration. I understand computational limitations and am only curious.

Also in terms of the computational setup, I wondered if the 3-hour spin-up was adequate for the considered cases? I would think there might be a substantial response and adjustment to the topography and land-use in the region at the considered scales when starting from scratch. Don't the results at the end when 48-hour runs were considered at least give some credence to this idea? To that end, which land-surface model was used and what is the terrain resolution? Research has also shown that long spin-up periods in the LSM are needed for improved land-surface representation. Given the focus on fluxes at the surface, those seem like relevant considerations, especially when the inner domain has a spacing of 750 m. That leads to questions about the PBL scheme. Was there any consideration to how running at 750 m might affect the PBL representation of a diurnal cycle given that this scale range is likely within the well-known grey zone of atmospheric turbulence?

Lastly, I would have expected more quantitative assessments of performance since this is a purported evaluation. For instance, many of the comparisons are presented as 2D color plots, which can be misleading. To be clear, I have no doubt v4fp1 was overall a better tool based on the authors' work, but I think some areas need more care. For instance, Section 3.2 introduces heat deficit as a proxy for cold pool strength, which is shown in Fig. 5 as a time trace. The positive traits of v4fp1 in the initial creation and decay periods were discussed at length in this section, but the middle portions to my eye show several periods where CTL and EXP were closer to the observations and yet the

text merely said "All model runs overestimated the heat deficit during the CAP period." I think it is reasonable to expect more explicit statistical analysis of this CAP period as it relates to the heat deficit and other fields given the title of the paper. As it stands, a lot of the evaluation is in the form of plots and many potential analyses are lost in, e.g., daily composites. Box plots are limited to fluxes, which are used as a proxy for clouds. Even then, there are plenty of locations where the mean biases in CTL and EXP are closer or as-close as in the v4fp1 cases. The results seem site dependent, which the authors address, but more discussion could be had as to how misses in certain periods or layers could result in better matches during other periods or in other locations (i.e., the so-called idea of getting the right answer for the wrong reason). That can all be related back to the issues above in terms of more discussion about the LSM, terrain, spin-up periods, grid spacing, etc.

Based on the above considerations, I believe this paper requires enough work that it would look substantially different than it does in its present form. In addition to these broad issues, I have a few specific issues that are listed below. Accordingly, I recommend that the manuscript require **major revisions** before it is suitable for publication in *EGUsphere*.

Specific Comments

Line 28 The authors reduced "cold-air pool" to "cold pool" on line 15, but use the full version here. Check for consistency.

Fig. 5 In print, these line colors were hard to delineate—especially the green ones. Is it possible to consider dotted lines for d02 domains of the same color as their d01 counterparts?

Fig. 6 There is an extra "Fig" in the caption text on line 2.

Line 275 Suggest changing "agreement in" to "agreement between."

Line 292 Consider rewording "with in general negative..."