Review of
‘The YOPP site Model Intercomparison Project (YOPPsiteMIP) phase 1: project overview and Arctic winter forecast evaluation’
by Day et al.

General comment :

This paper presents a numerical weather prediction (NWP) models intercomparison exercise that takes place in the framework of the Year Of Polar Prediction (YOPP) project. This exercice leverages a rich observational dataset of measurements collected at multiple Arctic and Antarctic sites during YOPP special observing periods to evaluate the ability of NWP models to predict temperature, wind and humidity in extreme polar conditions. The intercomparison of many models at many sites has been made possible through the design of a specific file format, the so-called Merged Data Files (MDFs) and Merged Model Data Files (MMDFs), and an associated processing chain in python. The sections 1 and 2 of the paper present all the sites and all the models involved while section 3 presents a first evaluation of models focusing on 7 Arctic sites where data are already available.

I am very impressed by how an international NWP community has collectively – and successfully - designed this ambitious, coordinated intercomparison exercise for polar regions. Collecting all the observations during the YOPP SOPs as well as model forecast from several NWP centers in a common format is already a fantastic achievement.
I am however less convinced by the content of the evaluation itself and by the conclusions drawn about model performances (especially sections 3.3 and 3.4). I think some further work is needed to really show how such a rich intercomparison and evaluation inform about possible shortcomings in the models’ physics and dynamics.
In summary, I really would like to see this paper published in GMD but I would also really appreciate the authors to strengthen some parts of the study before, following suggestions herebelow.

Major comments :

- Even though the authors mention that details on the sites are provided in Morris et al. (in prep) I think the reader does need some information about the landscape (distance from the coast, relief …) and terrain nature (vegetation cover, snow cover ..) at the different sites. I personally needed such kind of information at many places in the paper at many places in the paper (e.g., l255, l288, l440, l505-515, l569 …). A discussion on the representativity (or non-representativity) of station measurements with respect to the size of model meshes is also needed to disentangle actual model biases from model-observation differences inherent to possible very local nature of the measurements. I admit that adding such information implies increasing the length of the manuscript but a short description of the sites in this paper completed with a critical discussion on the spatial representativity is absolutely necessary to properly follow the analysis and understand the conclusions regarding models’ biases.

- The analysis of Figs. 6E-f (line 340-345) is not very conclusive. The authors leave the interpretation of the downward radiative flux biases for a future study but this aspect is essential to correctly understand the reasons behind the surface temperature biases. I would expect at least some additional analysis on the evaluation of the separate distributions of LWdn and SWdn and ideally some conditional analysis between cloudy and non-cloudy scenes. The idea behind this suggestion is to investigate whether models simulate the correct frequency of clouds and if the optical properties thereof is well reproduced.

- L370: I agree that LWdn + SWnet is the effective radiative forcing for the skin surface temperature (and indirectly to 2m temperature, this should be mentioned). Prior to investigate the response of the surface temperature, one first need to know if the albedo at the stations compares well with that observed at the sites (when available).

- L375: Is this due to the inability of models to simulate surface-atmosphere decoupling in clear-sky and windless conditions at those stations? Have you looked at the vertical profiles (simulations vs radiosonde) during these cases?

- L394-397: I do not fully agree here. In convective cases - the main driver of turbulent heat fluxes is indeed the convective instability at the surface driven by radiative forcing. However, in stratified (nocturnal) conditions the main driver of turbulence in the boundary layer (and of the sensible and latent heat fluxes) is the mechanical forcing i.e. the large scale wind speed (Van Hooijdonk et al. 2015, Van de Wiel et al. 2017, Vignon et al. 2017). All the subsequent sensitivity analysis in Sect. 3.4 is therefore incomplete and somewhat misleading for stable conditions. I would strongly recommend the author to carry out the study by separating convective cases from stable cases and to condition the analysis in stable conditions to certain large-scale wind speed classes (or to analyse the dependency of variables upon the large scale wind speed for different classes of LWdn+SWnet).

- Figure 13: In stable conditions, it has been shown that the turbulent heat flux increases then decreases with increasing stability, the maximum value separating a weakly stable from a very stable regime. This behavior is particularly well visible when conditioning the data to conditions with similar radiative forcing (Van Hooijdonk et al 2015). I would have been interested to see if the SHF data at Sodankyla show a clear maximum in stable conditions as well as comments on the ability of models to represent those stable boundary layer regimes (weakly stable cases in cloudy and/or windy conditions versus very stable regime in clear-sky windless conditions).

Minor comments :

- Table2: please specify that the timestep is the timestep of the physics (I guess).

- L255: Please recall the model-observation comparison period here.

- Figure 2 and 3: please indicate the local time at the beginning of the x-axes of the station to better identify daytime and nighttime in the graphs. A semi-transparent colour (gray?) shading in the figures themselved during the night periods may also help.

- Figure 5: Are statistics (interquartile ranges) calculated from model data at the same frequency as that of radiosounding?

- L471: Typo ‘Evaluation’

- Table 4: Roughness length can vary substantially depending on flow direction, snow cover … please specify the variability ranges as well.

- L535: What is ΔT?

- L546: I realize here that one has to know more specifically for each station which grid point(s) (with which ocean/land percentage) is considered for the evaluation. The information given at lines 141-142 is not sufficient to understand properly this paragraph.

- L557 ‘T is calculated using the temperatures observed at 18m and 32m so is not directly comparable with the models’ This sentence should be included in the main text I think.

- L580: ‘likely due to the single-layer representation of snow’: This is not shown in the paper, please remove the sentence or rephrase.

- L662: Please remove references to papers in preparation.

van Hooijdonk IGS, Donda JMM, Clercx JH, Bosveld FC, van de Wiel BJH (2015) Shear capacity as prognostic for nocturnal boundary layer regimes. J Atmos Sci 72:1518–1532

van de Wiel BJH, Vignon E, Baas P, van Hooijkdonk IGS, van der Linden SJA, van Hooft JA, Bosveld FC, de Roode SR, Moene AF, Genthon C (2017) Regime transitions in near-surface temperature inversions: a conceptual model. J Atmos Sci 74:1057–1073

Vignon E, van de Wiel BJH, van Hooijdonk IGS, Genthon C, van der Linden SJA, van Hooft JA, Baas P, Maurel W, Traullé O, Casasanta G (2017) Stable boundary layer regimes at dome C, Antarctica: observation and analysis. Q J R Meteorol Soc 143:1241–1253