MS No.: gmd-2018-330
Review, 11 March 2019

Title: The road weather mode RoadSurf driven by the HARMONIE-Climate regional climate model: evaluation over Finland

Authors: Erika Toivonen, Marjo Hippi, Hannele Korhonen, Ari Laaksonen, Markku Kangas, and Joni-Pekka Pietikäinen

Recommendation: [Major Revisions]

GENERAL COMMENTS:

This article presents an evaluation study across Finland of a 13-year long record of road surface temperatures and road ice, snow, and water storage parameters obtained with a road surface weather model driven by output from a regional climate model operated at 12.5 km resolution. The RCM is in turn forced at its lateral boundaries by atmospheric information from the ERA-Interim re-analysis. The emphasis of the analysis is on the performance of the road surface model compared with observations obtained at 25 road weather stations of which 11 were also equipped with optical sensors to establish the prevailing condition of the road surface.

Overall, the paper is coherently written, but in my opinion the scope is too much from the perspective of an NWP. The entire analysis assumes as if the modelling chain can be one-to-one compared with observations and a statistical machinery is applied resulting in skill scores which one usually sees in assessing the forecast performance of a prediction model. This approach does not match with the purpose of the study to evaluate a model system when operated in climate mode but still using observations (ERA-Interim) to constrain the large-scale model circulation to the observed synoptic-scale structure.

The next step, also mentioned by the authors, will be to run the RCM-RoadSurf modelling system driven by GCM output resulting from transient multi-annual simulation under prescribed emission scenarios. The biases then found will presumably be much larger than seen in this evaluation study, and any performance rating as if it were a prediction model will be deemed meaningless. The primary reason for that is huge biases in circulation and regime statistics in the GCM drivers compared to ERA-Interim. So, the authors can better focus on the role of circulation and regime drivers on the performance of their modelling chain, than focus on skill scores like RMSE and Pearson's correlation coefficients. Eventually, they want to draw credible conclusions how climate change information at the large scale will propagate through their RCM to the RoadSurf model.

In addition I would argue that the way the experiment has been set up makes it very difficult to conclude how the shortcomings in performance can be attributed to the model components that are used. Several times the authors mention that an issue might be related either to the warm and/or wet bias in their RCM or to features in the RoadSurf model. In that respect, I am wondering why the authors have not carried out a bias

adjustment to HCLIM-ALARO temperature and precipitation which serve as forcings to the RoadSurf model. Such an additional experiment would have a twofold benefit: a) to disentangle the bias in HCLIM from issues in RoadSurf, and b) to obtain a measure to what extent the biases in HCLIM affect the performance of RoadSurf. The latter would be very helpful in the analysis and interpretation of future GCM-driven experiments.

MAJOR COMMENTS:

1) I would strongly suggest to focus on the Finland area from the beginning. The discussion of the HCLIM-ALARO model performance for the whole of the Fenno-Scandian domain is distracting. There are always huge issues in the mountainous areas in Norway, for any RCM, and also in E-OBS, but they are not relevant for this study. Focus on Finland in Figs 3 and 5.

2) Do not only examine the bias in the monthly mean temperature, but also at a number of percentiles (e.g. P5,25,75,95). The diurnal amplitude in model temperature compared to observations is relevant here as well.

3) Similarly for precipitation. In addition to mean precipitation look at wet-day frequency (threshold 0.3 or 1.0 mm/day), and perhaps some exceedance percentiles. It provides much more insight than an RMSE score.

4) Can there be said anything about the accuracy of the RCM inputs other than near-surface temperature and precipitation that are used to drive the RoadSurf model.

5) As mentioned in the general comments it would be useful to apply a bias-adjustment on daily mean temperature and precipitation, also frequency of occurrence, to bring the HCLIM-ALARO temperature (e.g. quantile-quantile) and precipitation forcing in the same “statistical” ballpark as the observations.

6) As the RCM is operated at 12.5km resolution there should be reference to the efforts within EuroCordex in conducting evaluation (ERA-Interim driven) and transient (GCM driven) experiments at 12.5 km resolution across Europe with a variety of RCMs. For the evaluation study you best cite Kotlarski et al. (2014; doi: 10.5194/gmd-7-1297-2014).

7) Section 3.2.1 (“Road surface temperature”), after line 240 bothers me most. Why are all discrepancies blamed on the bias in temperature forcing, and not on potential issues with downwelling radiation, in particular biases in downwelling long wave radiation due to biases in cloud amount or cloud base.

8) Page 9, L260-266. The authors argue that the better skill obtained with the forcing from the NWP compared to this study can be ascribed to the higher resolution at which the NWP is operated. I tend to disagree on that, in my opinion the use of data-

assimilation when operating in NWP-mode will keep the model atmospheric state across the Finland region much closer to the observed state.

9) The statistical methods used in sections 3.2.2. are not suitable for evaluation purposes, they belong to the realm of NWP verification. I advise to take this section out or move it to the supplement.

10) The same applies to section 3.2.5 although I find the message (i.e. over-representing of storage of ice, under-representation of storage of water) quite useful. So I would advise to move the technical method to the supplement but keep the message in the main body of the manuscript.

OTHER COMMENTS:

1) It must be mentioned in the abstract that the HCLIM-ALARO simulation is driven by ERA-Interim
2) Abstract, L 13: remove "precisely"
3) Abstract, L 14, 18: replace "lack" by "absence" According to the text in Line 99 "the model does not take into account wintertime road maintenance operations …". From that line I conclude that there is no maintenance at all in the model. "Lack" may imply there is still some maintenance left. Please, adjust everywhere in the text, if needed.
4) Abstract, L 17: remove "simulated", it is already implied by "warm bias".

5) Introduction, L 24: "climate and weather information" →"weather and climate information".
6) Introduction, L34: "Finish temperatures …" →" Finish temperature records …"
7) Introduction, L42: replace "reliable" by "plausible" or "credible". It is not a prediction.
8) Introduction, L65: "13 year long simulations" → "13-year long simulations".

9) Page 3, L85-88: mention the source of the sea-surface boundary conditions (SST and sea-ice extent (probably also ERA-Interim)
10) Page 3, L92: "transfer in the ground …" →"transfer *into* the ground …"
11) Page 4, L95: ".. the elevation is taken into account …" The elevation of what or with respect to what?
12) Page 4, L107: "… we did not include any forecast periods". Suggest to add the phrase "implying that no in-situ observations are used to initialize and force RoadSurf."
13) Page 4, L119: Mention the version of the E-OBS dataset.
14) Page 4, L120: In addition to daily mean temperature, E-OBS also contains daily minimum/maximum temperature. Why not using these parameters for evaluation?
15) Page 4, L 122: remove "some"
16) Page 5, L 127: remove "some"
17) Page 6, L 180: the phrase "… such as from the possible biases in the input parameters ERA-Interim …" is confusing. Do you mean that land-surface information from ERA-Interim is used in forcing HCLIM-ALARO, or does this statement refer to the lateral/sea-surface boundary conditions specified from ERA-Interim?

18) Page 7, L197: “Similarly than in …” →”Similar to …”
19) Page 8, L227: “… during different months” →“… for different months”
20) Page 9, L268: “earlier” → “before”
21) Page 10, L296: “further” → “hence”
22) Page 10, L300-311: “the stations” → “stations” (about 11x)
23) Page 10, L307: “It could be expected …” →”It might be expected …”
24) Page 10, L316: “between the different stations” → “between stations”
25) Page 10, L317: “hypothesized” → “speculated”
26) Page 11, L343: “class occurred within a month” →”class *occurring* within a month”
27) Page 11, section 3.2.4 and Fig. 9 Perhaps you could briefly repeat that the road surface classes in the observations and the model do not entirely match.
28) Page 12, L356-357: rephrase last part of sentence as “i.e., the tendency *of the model* to underestimate frost and to overestimate ice with the same magnitude.”
29) Page 12, L360-361: “where much less maintenance” → ““where maintenance” and “is performed compared to …” →” is performed far less frequently compared to …”
30) Page, 12, L 362: “In real life,” → “In reality”
31) Page, 12, L 365-367: That is precisely the problem, because the bias in forcing temperature has not been adjusted the distinction between those two error sources cannot be made
32) Page, 12, L 375: No threshold used? Just, plainly 0 when the mean was 0?
33) Page, 12, L 379: “…storages might be slightly displaced or mistimed”. That is typical for NWP verification, but should not be relevant in an evaluation study.
34) Page, 14, L417-419: Second part of this sentence, “however …” is unclear. Please rephrase.
35) Conclusions, L 420-423. Like in the abstract it should be stated that HCLIM-ALARO is driven by ERA-Interim re-analyses.
36) Conclusions, L 422: “the skill of HCLIM- …” →”the skill of *the* HCLIM- …”
37) Conclusions, L 427: “undercath” →”undercatch”
38) Conclusions, L 427-428: “the modeled domain” →”the model domain”
39) Conclusions, L 432-433: Remove “However,”. Moreover, the absence of data-assimilation is most probably at least as relevant as the difference in horizontal resolution for explaining the poorer performance.

40) Conclusions, L 439: “This is of a great importance” →“This is of great importance”
41) Conclusions, L 439: “… are the most slippery…” →”… are most prone to slippery conditions …”
42) Conclusions, L442: “… than what the observations showed” → “ than is indicated by observations”
43) Conclusions, L447: “the 13 year long … period” → “the 13-year long … period”.

44) Figure caption 1: Does the displayed domain include or exclude the boundary relaxation zone? How wide is the zone in terms of grid points? The color “yellow” for Northern Finland is very hard to distinguish.