

Reviewer #2

We are grateful for the thought-provoking yet constructive comments, which have helped to greatly improve the manuscript. Our point-by-point response is below.

This manuscript presents OBSALL v1.0, an observation-space framework for monitoring and evaluating kilometer-scale Climate DT simulations using raw Earth observations. The topic falls within the scope of GMD, and the framework has the potential to complement existing climate-model evaluation approaches. However, several aspects of the manuscript's organization, methodology, interpretation, and discussion require clarification and further justification. I therefore recommend major revision before the manuscript can be considered for publication. My detailed comments are provided below.

Major Comments

1. Introduction: The Introduction contains the necessary background and motivation, but its current organization interrupts the progression of the argument. It first introduces Climate DT, AQUA, and OBSALL (lines 40–50), then presents the evaluation challenges associated with kilometer-scale simulations and advocates raw-observation evaluation (lines 51–63), before returning to conventional climate-model evaluation practices (lines 64–74). This is followed by the Vautard et al. example (lines 75–85), the OBSALL operational infrastructure (lines 86–98), possible future extensions (lines 99–104), and the paper outline (lines 105–110). The authors may consider reorganizing the Introduction to proceed from conventional evaluation practices, to their limitations for kilometer-scale simulations, to the potential value of raw observations relative to gridded products and reanalyses, and finally to OBSALL as the proposed solution. The section could conclude with a concise statement of the study's objectives and scope before presenting the paper outline. The discussion of possible future extensions to biogeochemical observations may fit more naturally in the Discussion. This organization would establish a clearer progression from existing practice, through the evaluation challenge, to the proposed framework and its demonstration.

Thank you for this comment which ended on having a strong influence on the revised manuscript. Introduction and Abstract are now re-written to closely follow Major comment #1 (except that the leading Paragraph expressing the scope is retained). The revision abides also to a similar comment by Reviewer #1 (who commented about the need to improve the literature review of current signal simulations approaches, such as COSP). Intro/Abstract now contain a review of traditional evaluation approaches, their limitations, the specific limitations related to km-scale evaluation, and how OBSALL remedies these limitations. This addition makes the discussion more balanced and respectful regarding earlier work in the field.

2. Section 2: Additional clarification is needed to help readers understand how the observation-space diagnostics are generated and interpreted.

o Lines 119–120: The manuscript states that observation-space evaluation “arguably provides higher fidelity than traditional model-space approaches.” Please clarify what “higher fidelity” means in this context and provide evidence supporting this statement. Because model fields are transformed from their native grids to the GSV/HEALPix representation before comparison with observations, it would be helpful to discuss the potential effects of this processing and, if possible, compare OBSALL results with conventional evaluation or native-grid sampling.

Thanks for this comment. The matter (“...fidelity...”) is now resolved (the 1st Paragraph of Section 2.1 is re-phrased). The relative merits of raw observations versus retrievals, etc., are now discussed in the Introduction, and Section 2.1 focuses what the subtitle indicates, the measurement equation.

The effect of the interpolation from native to HealPix grid is discussed in Section 2.2 in the original submission. The interpolation to HealPix contributes to observation error. However, for a fixed measurement station location, the nearest-neighbour grid column is always the same and does not add random variability, only a systematic influence. Also, the HealPix grid columns are interpolated run-time from the same native grid columns, again only a systematic influence due to the shift of grid column locations. These features are briefly touched in the ingress of Section 2 and in Section 2.2.

Observation operators are outside the simulation models in Climate DT and sample model state in HealPix format at run-time. It is impossible to repeat the multi-million core-hour simulations so that the operators would be inside the models, an implementation that does not exist. Therefore, we cannot compare our results with, e.g., native grid sampling.

o Lines 138–145: Please clarify the interpretation of the compound observation–model difference as an approximation of representativity error. Because the simulations are free-running and are not synchronized with the observations, these differences may contain contributions from multiple sources. The manuscript should distinguish more clearly between observation departures in data assimilation and statistical comparisons between observed and simulated climate distributions.

This distinction and all contributing factors are explained in a dedicated paragraph in Section 2.1 of the original submission (the 2nd full Paragraph after Equation 1). Therefore, no action is taken.

o Lines 164–184: Please provide additional technical information about the generation of observation-space counterparts, including the GSV resolution, native-grid-to-GSV interpolation, nearest-neighbor matching, variable transformations, and treatment of representativeness differences between stations and model grid cells. In particular, please explain how differences in elevation, land–sea classification, land cover, surface roughness, local terrain, and station exposure are evaluated or addressed. This information is important for understanding the methodology, reproducing the diagnostics, and assessing the sensitivity of the results to station-to-grid mismatch.

This is a complex request. GSV resolution is a matter of simulation protocol, subject to changes and beyond the control of OBSALL. Interpolation to HealPix is a system level issue and documented in detail by ECMWF/MPI-M/AWI in the context of streaming approach. The nearest-neighbour selection is explained in Section 2.2. Transformation from model quantities to observation-space quantities is signal-specific and implicit to operator H_T as explained in Section 2.2. Surface tiling is specific to each simulation model and it is not included in GSV by now (GSV is not a discretization for solving primitive equations, but rather a temporary storage for applications to consume atmosphere/ocean variables). There may be need to include more skin/surface variables in the future, for example to yield more accurate surface microwave and infrared emission modelling in remote sensing signal simulation. Local physiography is not included in GSV and the GSV grid columns inherit the local inhomogeneity influences of the parent model. $H(x)$ is computed at the nearest-neighbour grid column of GSV and the interpolation geometry from native-to-GSV-to-station location grid is the same throughout a multi-decadal simulation - there is nothing fishy here. There is a (small) systematic increase in observation error due to the interpolation to a “wrong” location (to GSV grid columns, not to columns at measurement stations, unless they happen to be exactly co-located), as explained in the original submission. Influence of the parent model physiography is not assessed here. The three example stations are selected so that the surrounding physiography is horizontally homogenous, i.e., representative of station conditions despite the few kilometres’ displacement of GSV column from the station. In the revised manuscript, the text is partly re-written and extended in the second last Paragraph of Section 2.2 to clarify this aspect (highlighted yellow).

3. Section 3: The section clearly describes the general concept of runtime monitoring, but additional information is needed to assess its operational implementation.

o Lines 240–253: Please clarify the criteria used to select the 541 stations, including the required record completeness over 1991–2014. For the observational quantile climatology, please specify the reference period, treatment of missing data, and procedure used to reintroduce the linear trend. The authors should also clarify the distinction between runtime processing of retrospective historical simulations and monitoring in a real-time setting.

The way in which the 541 stations are selected is documented in Appendix A.1. The methods used to conduct the quantile regression are described in Appendix A.2. Appendix A is added to the revised manuscript due to this, and other, comments.

o Lines 254–266: The manuscript states that significant deviations occurring simultaneously at many stations would raise an automated flag. Please document the statistical criteria and thresholds used to define these conditions. Because Figure 1 illustrates only a case that would not require intervention, an example showing the detection of a known simulation problem would more directly demonstrate the effectiveness of the monitoring system.

The text is intentionally conditional. The initial implementation raised flags far too frequently, and it remains unclear to us how to fine-tune the relationship between the misfit severity and the notification threshold. Although the mechanism exists, it is not currently in active use. No action is taken.

4. Sections 4 and 5: Additional clarification is needed regarding the statistical methodology, representativeness of the station-based evaluation, and interpretation of the results.

o Lines 282–345: Please provide complete temporal-sampling and data-completeness criteria for calculating monthly, seasonal, and annual statistics, as well as daily and annual extremes. In particular, please specify the minimum number and temporal distribution of observations required to define a valid day or year and explain how missing observations and incomplete heatwave or coldwave events are handled. Because the extremes are derived from observations available at 3-hourly intervals, the potential sensitivity of daily maxima, daily minima, and return values to temporal sampling should also be discussed. The term “diurnal range” should be clarified because the metric differs from the conventional daily maximum-minus-minimum range. For the heatwave and coldwave analysis, please also clarify how thresholds derived separately from each dataset affect the interpretation of model–observation comparisons.

Thank you for pointing out these issues, which need clarification. These questions led to substantial revisions and addition of a dedicated Appendix.

It is obvious that the absolute values of our statistics may be biased by missing data, even though missing data are relatively uncommon (Tables A.1–A.2). However, missing observations are not expected to have any major impact on the observation-minus-model differences because the times with missing observations are also masked from the model output (revisions are highlighted yellow, 2nd Paragraph of Section 4.1).

For the sake of simplicity, all the statistics were calculated when technically possible. Thus, in the worst possible case, a monthly mean might represent data for a single day, but the high degree of data availability makes this very unlikely. However, there is a four-UTC-time rule in the sampling of the diurnal cycle in the calculation of the monthly mean values, the diurnal range, the intramonth standard deviation, and the daily maxima and minima that are needed in the analysis of heatwaves and coldwaves (highlighted yellow, items 1, 2, 5 and 7 in the list of Section 4.1)

We believe that the original description of the diurnal range calculation was largely sufficient but have added one more clarification to motivate our methodological choice (highlighted yellow, note the addition “... sometimes generates a large difference between the daily maximum and minimum even during the high-latitude winter when solar radiation is negligible”).

The use of dataset specific temperature thresholds in the heatwave and coldwave analysis makes this analysis insensitive to biases in average maximum or minimum temperature. However, the number

and length of heatwave and coldwave events are affected by model-to-observation differences in the autocorrelation structure of temperature, and the magnitude indices (HWMI_e , CWMI_e) are additionally affected by the amplitude of temperature variability (highlighted yellow, end of Item 7 in the list in Section 4.1)

o Lines 325–347 and 413–424: Please provide further justification for the uncertainty estimates, including the assumption that simulated and observed variability are equal and the decision to neglect interannual autocorrelation. The manuscript should clarify whether the uncertainty ranges and map symbols indicate formal statistical significance or approximate sampling uncertainty. The short 24 year period, spatial dependence among stations, and use of a single realization from each model should be considered when interpreting trends and interannual variability. Please also clarify whether the trend differences can be distinguished from internal variability before they are interpreted as differences in the forced response.

Thanks for pointing out these needs for clarification.

Our choice for estimating the significance of model-to-observation differences from the observed variability alone was motivated by the needs of visualization. Significance thresholds also making use of simulated variability would be more precise, but they would make the presentation of sampling uncertainty in the seasonal cycle plots (e.g. Fig. 2) more complicated. This is because inter-model differences in the magnitude of variability would translate to model-dependent significance thresholds for the magnitude of model-to-observation differences, thus requiring three significance thresholds in the seasonal cycle plots instead of only one. In the revised manuscript, this is formulated as “The significance of the model-minus-observation differences could be estimated more precisely by also using information of the simulated variability, but this would complicate the visualization of the results.” (highlighted in yellow in Section 4.1).

Autocorrelation was neglected because it cannot be estimated with a reasonable accuracy from just 24 years of data. Nevertheless, we frankly admit that our uncertainty estimates are approximate only, and more likely to err to the non-conservative than the conservative direction. We also note that, in theory, initial condition ensembles of DestinE simulations would be helpful for better separating the effects of model biases from internal variability (highlighted yellow in the revised manuscript, last Paragraph of Section 4.1).

o Lines 357–365: Please provide additional support for the statement that the three selected stations are representative of their surrounding areas and clarify the criteria used to select these examples. The authors should also discuss how station-to-grid differences affect the extent to which results from point observations can be generalized spatially or interpreted as model errors.

The reason for selecting “representative” examples for this manuscript was to avoid distracting the readers’ attention with details, which steer the focus away from this first demonstration of the posterior evaluation methodology. In the selection of the example stations we have considered:

- (1) homogeneity of the surface. We selected such stations that are far away from large water bodies, so that such situation would not occur that the model grid point locates over water while the station is over land or vice versa. Prague and Zaragoza are airport stations, Sodankylä observatory is 1.7 km from an airport; outside of these airports the surface is considered homogeneous (Fig. 1 below).
- (2) surface orography. We selected such stations that do not have extreme elevation differences within 5 km radius from the station. In the case of these three stations, elevation differences within the 5 km radius are well below 200m. At Sodankylä, the elevation varies between 174m and 244m, at Prague between 240m and 400m, and at Zaragoza between 200m and 330m (Source: <https://en-gb.topographic-map.com/world/>). This means that the model grid point is in similar terrain and at similar elevation compared to the station.

- (3) sufficient observational coverage. We selected stations having good observational coverage in the period of 1991-2014 (Fig. 2 below).

Therefore, we argue that the posterior evaluation at the three selected stations is fairly unlikely to suffer from large station-to-grid discrepancies and can be generalized to the surroundings fairly well. However, this is not the case for all stations. As an example, we can show posterior evaluation at a few stations suffering of exceptional station-to-grid discrepancies (Fig. 3 below). At these stations it is clear that models struggle to represent the surface variables due to finite grid spacing. The models perform randomly better or worse depending on the proportions of land and sea in the nearest grid point in the native model grid.

We have modified the respective text slightly and added Appendix A.3 to better justify and express how the stations were selected and where the point observations are generalizable.

o Lines 503–510 and 548–553: Some conclusions attribute deficiencies in surface variability to specific model processes, including boundary-layer parameterizations, soil processes, and land–atmosphere coupling. Please provide additional process-oriented evidence for these attributions or present them more clearly as possible explanations. The statistical diagnostics identify differences between simulations and observations but may not, by themselves, establish their physical causes.

Thank you for this comment. As stated in the manuscript, within Climate DT, monitoring and evaluation information is fed into the continuous model development processes at the host institutions responsible for each simulation model. We do not claim expertise in all model formulations; nevertheless, we are able to provide valuable information on model behaviour in a multi-decadal simulation setting. Such multi-model simulations are only feasible through a large-scale initiative such as DestinE and offer significant value in identifying, for example, systematic model errors and biases. Investigating the underlying causes of these biases is beyond the scope of the present manuscript. In response to the comment, we have revised the text throughout to provide a clearer and more focused presentation.

5. Section 5: The Discussion would benefit from a more balanced discussion of the respective strengths and limitations of observation-space and model-space diagnostics. In particular, the authors should clarify which evaluation questions are best addressed by each approach and how OBSALL and AQUA complement one another within the Climate DT framework. Observation-space diagnostics retain the sampling and measured quantities of individual observing systems but are affected by observational coverage, temporal sampling, station-to-grid representativeness, observation-operator uncertainty, and the GSV transformation. Model-space diagnostics provide spatially complete information for evaluating circulation patterns, physical budgets, coupled processes, and regional-to-global structures, but may be affected by regridding and uncertainties in gridded reference products. Such a discussion would help reconcile the manuscript's description of OBSALL as complementary to existing approaches with statements suggesting that observation-space evaluation generally provides higher fidelity. The principal conclusion from this comparison could then be summarized briefly in Section 6.

Thank you for this comment. We fully agree and have acted accordingly. First, we have removed the apparent contradiction between presenting the approach as providing higher fidelity on the one hand and as being complementary on the other. As a result, the term "fidelity" and related expressions no longer appear in the manuscript. Second, and more importantly, Section 5 (Discussion) has been substantially rewritten in line with Reviewer #2's suggestion. This revision has resulted in a more coherent, organised, and fluent discussion compared with the original submission. Finally, as recommended, we have added a concise summary to Section 6 (Conclusions).

Minor Comments

- Line 358: The longitude of Sodankylä is reported as 26.619°N; this should presumably be 26.619°E.

Corrected

- The Prague station number is 115180 in the text (line 359) but 115810 in the Figure 2 caption (line 400).

115180 is correct. Corrected to caption of Figure 2.

- There is an inconsistency between the HadISD version numbers reported in Section 3 (v3.4.1.202403p, line 242) and Section 4 (v3.4.1.202401p, line 270). Please verify the version used and report it consistently.

v3.4.1.202403p is correct. Corrected to Section 4.

- Line 443: The text refers to “intramonth variability (Figure 6a),” but the correct reference appears to be Figure 7a.

Yes, the correct reference is Figure 7a. Corrected.

- Lines 482–483: The description of the ICON dew-point-difference bias is internally inconsistent. The manuscript first states that all three models overestimate the mean dew-point difference at approximately 60% of stations, but then states that the annual mean in ICON is too small at 83% of stations, with an average bias of -1.14°C . The latter statement also conflicts with Table 3, which reports an ICON annual-mean bias of $+1.21^{\circ}\text{C}$, with positive differences at 59% of stations. Based on the values in Table 3, the statement may instead refer to the diurnal range, for which ICON has a negative bias and only 17% of stations have positive differences. Please verify the intended statistic and ensure that the text and Table 3 report consistent values.

We are sorry for the confusion. As suggested, the last sentence inadvertently mixed the mean value with the diurnal range of the dew point difference in ICON. The corrected text (highlighted yellow) now reads:

“Globally, all three models overestimate the dew point difference at about 60% of the stations, with an average bias of approximately 1°C (Table 3 and Fig. S18). In ICON, however, the annually averaged diurnal range of the dew point difference is too small at 83% of all stations, with an average bias of -1.17°C .”

- Lines 107–108: The Introduction states that Section 4 focuses on the simulation of European climate. However, Section 4 presents numerous global station maps and discusses results across the global station network. Please clarify the intended geographical scope and describe it consistently throughout the manuscript.

Thanks. Introduction and Section 4.2 now make the point (“The focus here is on observed versus simulated climate in Europe at three subjectively selected surface stations and on global statistics”).

- Table 3: The final otherwise empty row, which repeats only the column headings, appears to be redundant.

This is intentional, as it improves the readability of the lower part of the table, and in the case of GMD, paper usage is not a concern. However, if preferred, the Technical Editor may remove the line.

- The code and data section still contains the placeholder “[to-be-specified],” which must be resolved before publication.

Thanks, this is now added.

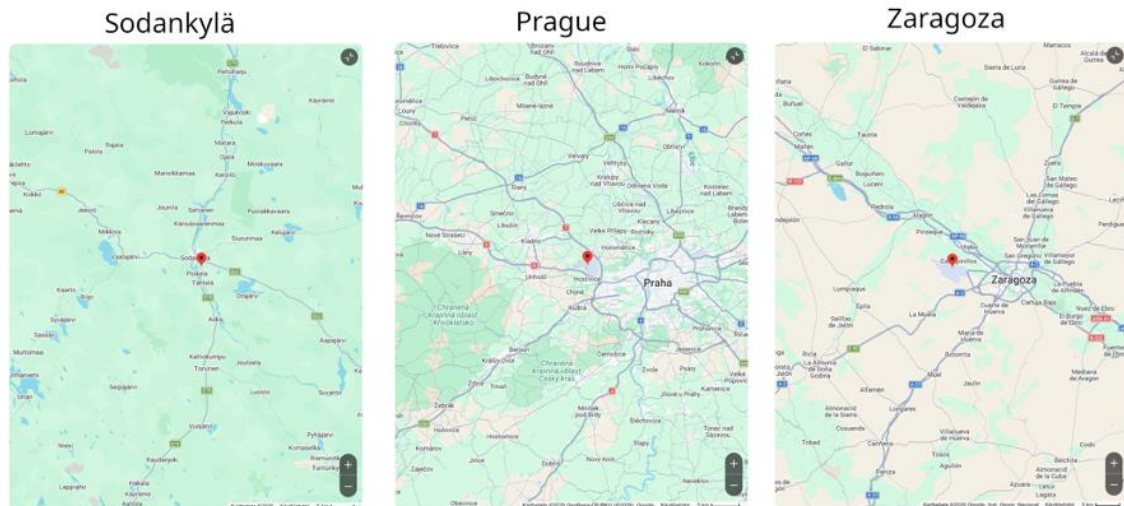


Figure 1. Geographical maps describing the surroundings of the selected stations marked with red dots. Green is forest, brown is open terrain (grassland, agricultural land, etc.), blue is water and shades of grey are urban areas. (Source: Google Maps).

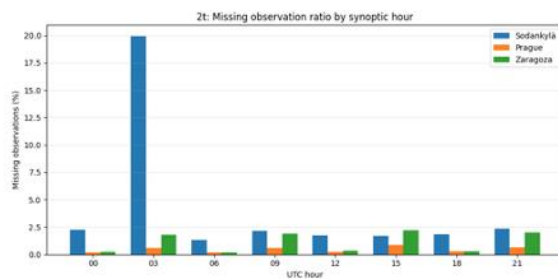


Figure 2. Percentage of missing 2-metre temperature observations within the time range of 1 Jan 1991 to 31 Dec 2014 for each UTC hour for the three stations. Statistics for 2-metre dewpoint depression and 10-metre wind speed are visually similar.

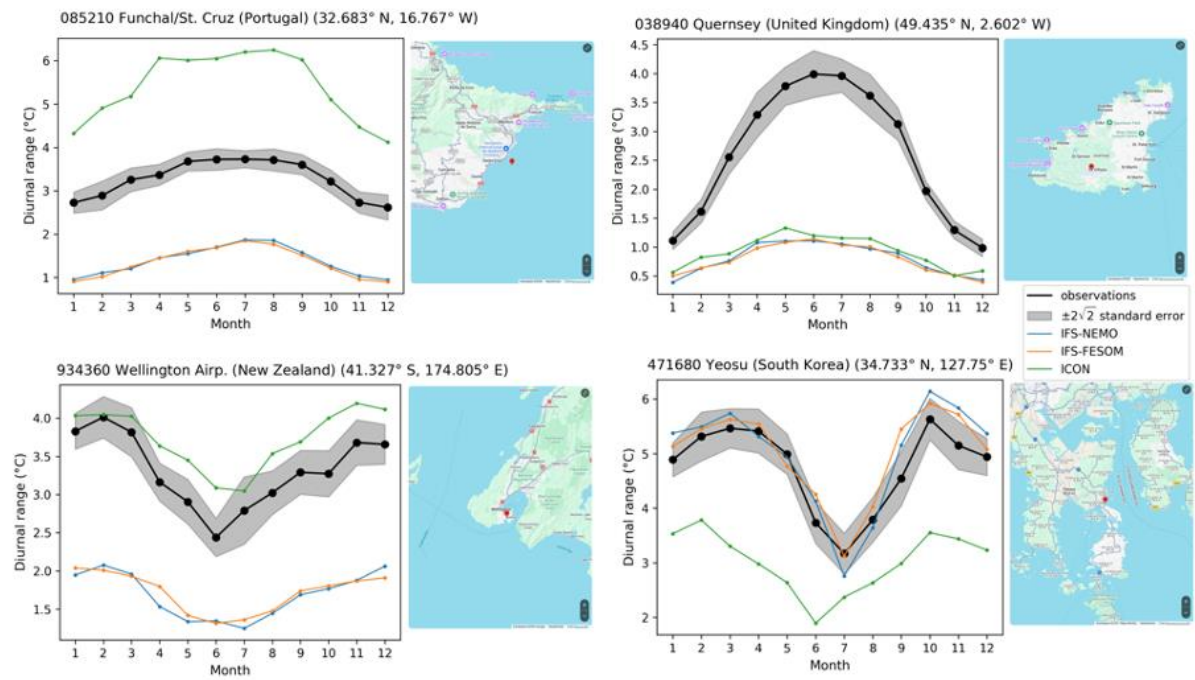


Figure 3. Examples of 2-metre temperature diurnal range at stations that are difficult to be represented by models. (Source for the maps: Google Maps).