Reviewer Comments on "Observation-based evaluation of the Destination Earth climate change adaptation digital twin simulations using OBSALL v1.0"

By Heikki Järvinen, Jouni Räisänen, Lauri Tuppi, Clément Bouvier, Antonio Sanchez-Benitez, Juniper Tyree, Antti Toropainen, Paolo Davini, Francisco Doblas-Reyes, Thomas Jung, Daniel Klocke, Jenni Kontkanen, Sebastian Milinski, Matteo Nurisso, Himansu Kesari Pradhan, Irina Sandu

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

- **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.

3. **Section 3:** The section clearly describes the general concept of runtime monitoring, but additional information is needed to assess its operational implementation.
   - **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.
   - **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.
4. **Sections 4 and 5:** Additional clarification is needed regarding the statistical methodology, representativeness of the station-based evaluation, and interpretation of the results.
   - **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.
   - **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.
   - **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.

- **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.

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.

Minor Comments

- Line 358: The longitude of Sodankylä is reported as 26.619°N; this should presumably be 26.619°E.
- The Prague station number is 115180 in the text (line 359) but 115810 in the Figure 2 caption (line 400).
- 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.
- Line 443: The text refers to "intramonth variability (Figure 6a)," but the correct reference appears to be Figure 7a.
- 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 °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.
- 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.
- Table 3: The final otherwise empty row, which repeats only the column headings, appears to be redundant.
- The code and data section still contains the placeholder "[to-be-specified]," which must be resolved before publication.