https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

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

Heikki Järvinen[1], Jouni Räisänen[1], Lauri Tuppi[1], Clément Bouvier[1], Antonio Sanchez-Benitez[2], Juniper Tyree[1], Antti Toropainen[1], Paolo Davini[4], Francisco Doblas-Reyes[6], Thomas Jung[2,3], Daniel Klocke[7], Jenni Kontkanen[8], Sebastian Milinski[9], Matteo Nurisso[5], Himansu Kesari Pradhan[2], Irina Sandu[9]

[1]Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, Helsinki, Finland
[2]Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung (AWI), Bremerhaven, Germany
[3]Institute of Environmental Physics, University of Bremen, Bremen, Germany
[4]Consiglio Nazionale delle Ricerche, Istituto di Scienze dell'Atmosfera e del Clima (CNR-ISAC), Torino, Italy
[5]Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Torino, Italy
[6]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[7]Max-Planck-Institut für Meteorologie (MPI-M), Hamburg, Germany
[8]CSC – IT Center for Science, Espoo, Finland
[9]European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany

Correspondence to: Heikki Järvinen (heikki.j.jarvinen@helsinki.fi)

Abstract. Destination Earth Climate Change Adaptation Digital Twin (Climate DT) is setting up an operational simulation framework that enables to produce bespoke climate and impact-sector information, supporting evidence-based decision-making and fortified societal resilience. Climate DT combines global kilometer-scale climate models and a range on impact-sector applications in a unified operational workflow, producing both multi-decadal climate projections and storyline simulations at the scales of the impacts of climate change and extreme events. This imposes new demands for the evaluation of simulation quality. The question is, which reference data are adequate to evaluate the rich process-level variability present in these new-generation models. Here, we advocate the use of raw Earth observations for assessing physical fidelity of the modelled fine-scale variability, operating exclusively in the observation-space.

The Climate DT framework (https://platform.destine.eu/) enables unique evaluation of the simulation quality using Earth observations directly, as the model output data is available run-time in a near-native grid. This is in stark contrast to traditional model-space evaluation using archived simulations and gridded reference data. Climate DT simulations are evaluated in observation-space prior to any spatio-temporal truncation, exposing their process-level variability for inter-comparison with raw Earth observations and bridging the gap between modelling and observing the Earth system. The synergy aspect here is the extensive sharing of observation modelling infrastructure with data assimilation in numerical weather prediction.

This article presents the Climate DT concept to assess simulation quality with Earth observations and showcase it through the lens of synoptic surface observations. The evaluation covers the simulated mean, trend, variability, and extremes of historic simulations from 1991 to 2014 of the IFS-NEMO, IFS-FESOM, and ICON models. Annual cycles of 2-metre temperature, and (to a somewhat lesser extent) humidity and 10-metre wind speed are generally well-simulated. However, these quantities contain also significant process-level weaknesses, such as too weak process level variability at diurnal, intramonth, and

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

interannual time-scales. The evaluation results indicate that while there is still work to be done to improve the Climate DT simulation models, it is a novelty that they can now be examined at such level of detail and objectively interfaced with raw Earth observations containing the corresponding process-level imprints. Furthermore, since climate change adaptation mostly occurs at local level, the raw observation-based evaluation informs directly the scales of interest.

1 Introduction

The Climate Change Adaptation Digital Twin (Climate DT), implemented in the Destination Earth initiative of the European Commission, is designed as an operational simulation framework, where state-of-the-art coupled atmosphere-ocean climate models are run operationally to feed user-oriented applications and support informed decision-making and societal resilience (Sandu 2024; Wedi et al. 2025; Doblas-Reyes et al. 2026; https://destine.ecmwf.int/climate-digital-twin/, last access 20 May 2026). The reliability of this framework is underpinned by the high quality of the simulation data it generates (Dee et al. 2024). Two separate systems are designed for monitoring and evaluating Climate DT simulation data quality: AQUA (Nurisso et al. 2025, 2026), using gridded reference data, and OBSALL, using raw Earth observations. In a nutshell, AQUA assesses climatological realism whereas OBSALL assesses process realism of the Climate DT simulations. The focus of this article is on OBSALL v1.0, the observation-based monitoring and evaluation system, deeply rooted in numerical weather prediction and statistical climate model validation. Examples illustrate the solution to operate with observation-space quantities and its performance in Climate DT model evaluation as viewed through the lens of selected in-situ surface observations.

Climate DT operates three storm-resolving kilometre-scale atmosphere-ocean models in multi-decadal time-scale, resolving ever finer details of the motion spectrum. In the Phase 2 of Climate DT (May 2024-April 2026), the resolution of these models is 5 km in the atmosphere and 5-9 km in the ocean. This imposes new requirements for the evaluation of simulation quality. The question here is, which reference data are adequate to evaluate the rich process-level variability present in these new-generation models. For example, the nominal resolution of ERA5 (Hersbach et al. 2020) is 32 km, with the analysis increments computed at the resolution of 62 km. These data are well-suited for evaluating realism of any large-scale aspect of mean climate, general circulation, or synoptic variability of the Climate DT simulations. At process-level, there is, however, a significant resolution gap such that the state-of-the-art global reanalyses do not provide reference data to evaluate realism of the highest-end of spatio-temporal thermodynamical and motion spectra. Furthermore, climate-relevant decisions are often taken at local (city) level and hence, it is advisable to carry out point-wise evaluation prior to any spatio-temporal truncation of model data and thus ensure faithful inter-comparison of observed versus modelled variability. In this paper, we advocate the use of raw Earth observation-based approaches to complement existing evaluation methods and lay the foundation for deploying observation modelling to interface climate models consistently and objectively with observations.

Earth observations are commonly used in climate model evaluation. In typical posterior evaluation, archived simulation data is inter-compared with observation-based information, which is presented in model-space geophysical quantities. Simulation data is archived at reduced spatial and temporal resolution, such as in the successive phases of Coupled Model Intercomparison

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

Projects (CMIP; Eyring et al. 2016; https://wcrp-cmip.org/cmip-data-access/, last access 20 May 2026), enabling efficient access and data-distribution solutions. As an example, CMIPs operate within this framework and enable observation-based quantification of, for instance, progress across different CMIP phases (Bock et al. 2020) using either point-based (e.g., Durre et al. 2006) or gridded in situ observations (e.g., Frick et al. 2014), remote sensing products (e.g., Stocker et al. 2018; Moreno-Chamarro et al. 2022), and reanalysis data (e.g., Hersbach et al. 2020). These reference materials also serve as input for dedicated validation tools (e.g., Eyring et al. 2020; Zhang et al. 2022; Nurisso et al. 2026). Inter-comparisons can include preprocessing steps, such as re-gridding of simulation data into a common grid to evaluate models using quality-controlled surface station observations (e.g., Shen et al. 2022).

Importance of Earth observations in detecting decadal-scale climate change signals is underscored by Vautard et al. (2010), who reported a long-term decline in near-surface wind speeds, particularly over the Asian continental region. Closer examination identified the most plausible explanation to be a gradual increase in surface roughness driven by expanding vegetation. This environmental change leaves an imprint in surface station observation records, which are sensitive to surface fluxes of momentum, heat, and water. However, this decadal-scale signal is subtle and at the time of the study, it had not yet been fully captured by common climate reanalyses. The results of Vautard et al. (2010) therefore reinforce the case for raw observation-based evaluation of decadal climate simulations. This illustrates a broader limitation of reanalyses and gridded products: they can mask or filter process-level signals that are directly visible in station records. Kilometre-scale models resolve variability that operates on similar spatial scales as station representativity errors; therefore, evaluation against raw observations – rather than heavily processed gridded products – is essential to assess whether these models capture physically meaningful variability or merely generate small-scale noise.

The Climate DT workflow solution constitutes as step change compared to the usual climate projection approaches by pushing climate models to km-scale solution on pre-exascale EuroHPC supercomputers. This is done by leveraging decades-long experience on state-of-the-art climate modelling and operational numerical weather prediction (NWP) in order to operationalise production of multi-decadal-scale climate simulations (Bauer et al. 2021; Jacob et al. 2023; Stevens 2024; Wedi et al. 2025). In Earth observation-based evaluation, the key enabling technology is the seamless and homogenous data access, as detailed in Doblas-Reyes et al. (2026). Regardless of which simulation model is running, the model state vector is streamed into a dense model-independent grid, termed “generic state vector” (GSV), enabling time-critical data access while the main simulation workflow is active. Observation models (or, “operators”) consume GSV data in parallel tasks, formulated as a time-critical one-pass algorithm (OPA; Grayson et al. 2025), and compute observation-space quantities to be stored in observation data base (ODB). This approach has important commonalities with NWP data assimilation, where accurate and comprehensive observation modelling is only possible using full-resolution model data while all essential state vector elements are still accessible, such as surface emissivity – essential for many remote sensing measurements but not for most climate studies.

With the release of OBSALL v1.0, evaluation of the Climate DT simulation data focuses on the use of raw Earth observations. Presently, Climate DT deploys three coupled atmosphere-ocean models for multi-decadal and storyline simulations – ICON, IFS-NEMO, and IFS-FESOM. These models focus on the physical aspects of the climate system, and the evaluation relies on

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

Earth observations and observation modelling tools commonly applied in NWP. If the selection of Climate DT models will in the future expand towards including, e.g., biogeochemical cycles, the observation-based evaluation concept can be extended to cover the associated measurements, such as those from the Integrated carbon observation system (https://www.icos-cp.eu/data-services/about-data-portal/, last access 20 May 2026).

The outline of the paper is as follows. Section 2 presents the method to compute observation-space counterpart from model variables, the simulation and observation data. Section 3 focuses on run-time (online) monitoring of ongoing simulations. Section 4 presents the statistical methods applied in posterior evaluation and the main findings regarding the current Climate DT models, focusing on simulation of European climate. The evaluation covers the historic simulations from 1991 to 2014 with the statistics for mean climate and its trend, as well as annual and diurnal cycles, irregular variability, and extremes. Discussion (Section 5) and Conclusions (Section 6) conclude the paper. The Supplement contains a wider selection of Figures.

2 Data and methods

Observation modelling is commonly applied in data assimilation of numerical weather prediction (NWP), while its systematic application is rare in climate modelling. Therefore, we start by introducing its theoretical basis, the measurement equation (Section 2.1) and the computation of observation-space quantities from model-space variables (a.k.a. observation projection; Section 2.2).

2.1 Measurement equation

In traditional approaches to climate model evaluation, model output is left in its native form while reference data are transformed into model-space quantities, enabling direct intercomparison (reference–minus–model; e.g., Gampe et al. 2019). The Climate DT framework, by contrast, supports observation-space evaluation, which arguably provides higher fidelity than traditional model-space approaches. Here, observations remain intact, and model counterparts in observation-space are computed for each observation type using dedicated observation operators, enabling observation-space intercomparison (observation–minus–model-counterpart). This observation-space formalism is standard in state estimation, such as data assimilation in NWP, where the measurement equation specifies how observed quantities relate to the underlying model state (e.g., Kaipio and Somersalo 2005; Kalnay 2002). It is written as

$$\boldsymbol{y}_k = H(\boldsymbol{x}_k) + \boldsymbol{\varepsilon}_k, \quad (1)$$

where $\boldsymbol{y}_k$ is the observation vector at time k, $\boldsymbol{x}_k$ the model state, H the (possibly non-linear) observation operator, and $\boldsymbol{\varepsilon}_k$ the observation error. In NWP, $\boldsymbol{\varepsilon}_k$ is typically small because the perpetual cycling of model time evolution and observation updates keeps the state $\boldsymbol{x}_k$ close to the observations $\boldsymbol{y}_k$. The observation error $\boldsymbol{\varepsilon}_k$ comprises three components: instrumental error arising from inaccuracies in the measurement process ($\boldsymbol{\varepsilon}_{o,k}$), observation-modelling error ($\boldsymbol{\varepsilon}_{H,k}$) due to imperfections in the operator H, and representativity error ($\boldsymbol{\varepsilon}_{r,k}$) resulting from the inability of the discrete state vector $\boldsymbol{x}_k$ to fully represent the spatio-temporally continuous physical system.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

There is no practical method to separate the error components ($\boldsymbol{\varepsilon}_{o,k}$, $\boldsymbol{\varepsilon}_{H,k}$, $\boldsymbol{\varepsilon}_{r,k}$) from one another. They nevertheless help to conceptualise the composition of the compound observation error $\boldsymbol{\varepsilon}_k$. In addition to instrument noise, $\boldsymbol{\varepsilon}_{o,k}$ can contain gross errors such as due to the analog-digital conversion and instrument freezing of in-situ observations. $\boldsymbol{\varepsilon}_{H,k}$ contributes to random and systematic errors and instrument-specific gross errors, such as a failure in detection of cloud contamination of passive remote sensing instruments. $\boldsymbol{\varepsilon}_{r,k}$ collectively point to modelling inabilities, for example, a hydrostatically-constrained model state in representing non-hydrostatic phenomena.

In NWP data assimilation (Lean et al. 2021), the compound observation error $\boldsymbol{\varepsilon}_k$ is a small residual in Eq. (1). In the free simulations of Climate DT, it is relatively much larger. The reasons are as follows. While the instrumental error ($\boldsymbol{\varepsilon}_{o,k}$) is the same in both cases, the observation modelling error ($\boldsymbol{\varepsilon}_{H,k}$) is larger in Climate DT due to the additional interpolation step of the native output datacubes to the GSV format. The representativity error ($\boldsymbol{\varepsilon}_{r,k}$) is relatively speaking very large in Climate DT because interannual and synoptic variabilities are out-of-sync in simulations and observations. Its order of magnitude equals the difference between two arbitrary states that are projected into the observation space. In summary, the inequality $\boldsymbol{\varepsilon}_{o,k} + \boldsymbol{\varepsilon}_{H,k} \ll \boldsymbol{\varepsilon}_{r,k}$ mostly holds in Climate DT and hence, the compound observation error $\boldsymbol{\varepsilon}_k$ can be considered as a good approximation for the representativity error $\boldsymbol{\varepsilon}_{r,k}$, which we would ideally like to study, if it was not for the inseparability of the error components.

Instrumental error ($\boldsymbol{\varepsilon}_{o,k}$) includes uninformative gross errors, which observational quality control (QC) aims to detect and remove. Here, a quality-controlled set of synoptic surface observations (Dunn et al. 2012, 2014, 2016) is used. A number of automated QC tests have been applied to the data set to ensure, e.g., internal consistency of station time-series (logical cross-checks between variables from one time instant), climatological consistency (removal of duplicates and climatological outliers, ensuring realism of diurnal cycle), and geographical consistency (nearest-neighbour check). This set of QC tests is imaginative and leads to a high-quality, trustworthy data set. An alternative would be to use quality-controlled observations from reanalyses. Gross errors are searched, for example in ERA5, in the background check using the current state estimate as a reference, which automatically accounts for annual and diurnal cycles and effects due to local physiography. The analysis check further ensures consistency with surrounding observational information. These two steps (Andersson and Järvinen 1999) correspond approximately to the procedure of Dunn et al. (2012, 2014, 2016). Therefore, observations deployed in the Copernicus reanalyses (Hersbach et al. 2020), together with the QC feedback information, would form a comprehensive observational data set for Climate DT. This future option is, in fact, becoming relevant with the move of Climate DT towards (near) real-time capabilities involving storyline simulations and their continuous observation-based evaluation (Shepherd et al. 2018; Sánchez-Benítez et al. 2022; John et al. 2026).

2.2 Computation of observation-space quantities

The measurement equation (Eq. 1) contains the observation projection to compute observation-space quantities from model-space variables, the so-called model counterparts in the observation space (Pailleux 1990). It is often decomposed as follows:

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

$$H(\boldsymbol{x}_k) = H_R\left(H_T\left(H_I(\boldsymbol{x}_k)\right)\right) \quad (2)$$

where H_I is the interpolation of the model state $\boldsymbol{x}_k$ to observation location, including possible spectral transforms, H_T the transformation from the model state-space variables to the observation-space quantities, and H_R the application of instrument-specific response functions.

The implementation of operator H in Climate DT is as follows. For interpolations (H_I), OBSALL applies ECMWF's Polytope service (Leuridan et al. 2025), which enables to access the DestinE digital twin data at native resolution, and has full backward compatibility with ECMWF's Meteorological Archival and Retrieval System (MARS; Raoult 1997; https://confluence.ecmwf.int/display/UDOC/MARS+user+documentation/, last access 20 May 2026). Polytope has the capability of extracting arbitrary geometrical shapes from data hypercubes. Presently, it is used to extract vertical GSV profiles at observation locations but keeps options open for example to slant profile extraction. The interpolation from the model-specific native grid to the model-independent GSV representation – a grid termed HealPix with grid cells equal in area but complex in shape (Górski et al. 2005) – is considered here a part of H_I. Interpolation from the HealPix grid columns to the observation locations is currently disabled in favour of selecting the nearest-neighbour. This aims to minimize confusion in process-level signatures and surface influences in addition to what may already take place in the interpolation to GSV. Regarding H_T and H_R, in situ observations are mostly of the same physical quality as the model variables with a small exception in humidity variables (specific versus relative humidity). Also, in radio-sounding measurements, horizontal wind represents mean value in a layer, where the sonde drift is monitored. This part of observation modelling is omitted here.

Finally, OBSALL contains a remote sensing data demonstrator in the form of AMSU-A radiances from NOAA/EUMETSAT. The AMSU-A operator $H_R(H_T(\cdot))$ is the NWP SAF Radiance Simulator v3.2 (https://nwp-saf.eumetsat.int/site/radsim-v3-2-released/, last access 20 May 2026). Accuracy of AMSU-A modelling using the Climate DT GSV files as input benefits from the high vertical resolution of the GSV data hypercube. Presently, the measurement data cover the years 2010-2020. Our early results indicate that AMSU-A radiances fit well to the online monitoring concept of Climate DT and can be used in parallel to other Earth observations (detailed results will be published separately).

2.3 Workflow management

The OBSALL data processing workflow is implemented within the integrated workflow of Climate DT. This Autosubmit-based workflow manager (Manubens-Gil et al. 2016) intermittently triggers the OBSALL-specific workflow, which is analogous to the so-called trajectory run in four-dimensional variational data assimilation (4D-Var; Rabier et al. 2000). In a 4D-Var outer-loop, high-resolution simulation over the assimilation window of, e.g., 12-hours applies observation operators at appropriate times of available observations and stores the model counterparts $H(\boldsymbol{x}_k)$ for cost function evaluation.

The OBSALL-specific workflow is conceptually similar to the description above, with the main difference that the simulation is free-running over multi-decadal time scales and therefore has no synoptic correspondence between the simulated state and the observations. In this configuration, the cost-function computation is replaced by the calculation of evaluation statistics.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

OBSALL shares several infrastructures with ECMWF and, more broadly, with NWP systems. The observation operators are identical to those used in NWP data assimilation (https://www.ecmwf.int/en/research/data-assimilation/observations/, last access 20 May 2026); https://www.ecmwf.int/en/elibrary/81623-ifs-documentation-cy49r1-part-i-observations/, last access 20 May 2026). Data-extraction codes are shared through the Polytope service (Leuridan et al. 2025). The relational observation database (ODB; Fouilloux 2009) and its Python interface (pyodc; Hodson 2025) are likewise common components. These tools are employed in Climate DT in their standard configuration, without any modifications.

Individual Climate DT simulations are termed here as experiments. At the start-up of an experiment, the ODB files only contain the observations: synoptic surface station data (Dunn et al. 2012, 2014, 2016), upper-air soundings (Madonna et al. 2022), and AMSU-A radiances of NOAA/EUMETSAT. ODB is first augmented with pre-computed daily quantiles of the observed climatology, which are needed in online monitoring. Next, the list of station coordinates is collected from the ODB for the first GSV data extraction. As the simulation begins and the first GSV datacube becomes accessible, profiles of GSV data at each observation location are extracted, observation projection is computed, and the ODB files are augmented with model counterparts. In parallel, the next station list is collected before the time-critical GSV for the subsequent time-slot becomes available. After each completed model year, annual online monitoring statistics and figures are generated. After the climate simulation is successfully finished in the main workflow and the ODB files are complete, the statistics and figures for the posterior evaluation are generated. The complete experiment-specific fully-augmented ODB is archived as a full-resolution trace of the simulation in observation space. All evaluation statistics and figures are stored into a dedicated repository, termed Dashboard, where they are accessible, for example, for model development purposes.

Observations and their model counterparts in observation space are long-term archived in ODB format and can be found based on an experiment-specific identifier. As Climate DT progresses and the number of simulations increases, the library of ODB files expands. Simulations can thereby be accessed both in gridded format as GSV files and in observation-space format as ODB files. This allows, for instance, assessing the magnitude of systematic errors in different model versions and generations. The computational throughput of IFS-NEMO on EuroHPC LUMI-C/G is tuned to approximately one simulated year per day (SYPD). At this rate, GSV files are generated roughly every ten seconds. Accordingly, the computational resources allocated for observation projection and ODB updates must be provisioned to operate at this cadence. In practice, the integrated workflow manager reserves the required parallel processing resources when launching an OBSALL batch job, ensuring that each newly completed GSV file is processed within these bounds – effectively implementing a time-critical, one-pass algorithm.

2.4 Simulation data

The multi-decadal simulation data in this paper (Table 1) cover three simulation models at 5 km horizontal resolution in the atmosphere: IFS-NEMO, IFS-FESOM, and ICON. The historic simulations are from the Climate DT Phase 2, generated coherently in operational environment early in 2025 (designated o25-1). The common simulation period in the observed past

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

covers years 1991-2014, on which the observation-based posterior evaluation will focus. The Phase 2 multi-decadal simulations covering the period 1990-2050 with the three Climate DT models are now available via the DestinE platform.

Table 1. Technical details of the model versions used in evaluation of Destination Earth Climate DT simulations, providing the input for OBSALL.

Climate DT model	IFS-NEMO	IFS-FESOM	ICON
Simulation	o25-1; historical period 1991-2014 (o25-1 = operational, batch one of 2025)		
Atmospheric model	Integrated Forecasting System (IFS) cycle 48r1; DestinE suffix climateDT_20250521	IFS cycle 48r1; DestinE suffix climateDT_20250521	ICosahedral nON-hydrostatic Atmosphere (ICON-A); DestinE Phase 2 v1.2.0
Ocean model	Nucleus for European Modelling of the Ocean (NEMO) v.4.0.7	Finite-volumE Sea ice-Ocean Model (FESOM) v.2.6.5.1	ICON-Ocean; DestinE Phase 2 v1.2.0
Atmospheric resolution	Tco2559 (~5 km) with 137 vertical levels	Tco2559 (~5 km) with 137 vertical levels	R2B9 (~5 km) with 90 vertical levels
Oceanic resolution	eORCA12 (~9 km) with 75 vertical levels (25 levels in the upper 100 m)	NG5, unstructured triangular mesh (~5 km)	R2B9 (~5 km)
References	Hohenegger et al. (2023); Rackow et al. (2025); Segura et al. (2025); Wedi et al. (2025); https://platform.destine.eu/de/services/documents-and-api/doc/?service_name=climate-dt-user-guide/, last access 20 May 2026; https://www.icon-model.org/, last access 20 May 2026;		

3 Run-time monitoring of ongoing simulations

The Climate DT monitoring and evaluation system is primarily designed for quality assurance in support of producing user-oriented adaptation information. In run-time (online) monitoring, the goal is to enable "on-the-fly" early-detection of potential issues that could degrade simulation quality – such as mis-specified boundary files – and thereby allow early-rejection of problematic simulations early in the process by operators of the Climate DT workflow system. In posterior evaluation, in contrast, the objective is in detection of systematic model errors relative to observations. This supports the Climate DT uncertainty quantification and can enhance usability of the Climate DT adaptation information. It also feeds into the continuous model development process, in liaison with the modelling teams.

Here we focus on synoptic in-situ surface observations (SYNOP) from stations that report consistently over the common simulation period 1991-2014 (Table 1). The observations are extracted from the Hadley Centre Integrated Surface Database Centre (HadISD), version 3.4.1.202403p (Dunn et al. 2012, 2014, 2016). The selected subset of 541 stations reports at least four times a day the basic meteorological variables: surface pressure p_s, 2-metre temperature T and dew point T_d, and 10-metre wind components u and v. Since the generic state vector (GSV) is available each hour, the projection is carried out hourly, too. However, in comparison with observations, only projections from observation times are included in the data analysis. As an illustration, 541 stations and five variables lead in total to 64920 individual projections each simulation day.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

The key output of OBSALL online monitoring is visual information about the performance of climate simulations. The monitoring is based on comparison of simulated values to quantile climatology derived from observational records. The quantiles are constructed separately for each variable at every station for the main synoptic observing times (00, 06, 12, 18 UTC). The quantile climatology is constructed by first subtracting linear trend from the observational time series, and then performing a quantile regression to the detrended data by fitting the first four annual Fourier components to all quantiles. Lastly, the linear trend is reintroduced in order to account for the climate change present in observational time series. These precomputations are stored in ODB for easy access.

An example of online monitoring is shown in Figure 1 for the simulated 2-metre temperature at 00 UTC at the Sodankylä observatory (IFS-NEMO, model year 2014). The quantile plot (left panel) offers an intuitive way to view the simulation within the observed climate and decide whether it warrants closer inspection. In this case, there is a slight cold bias in spring/summer in the simulation. The corresponding ranked percentile histogram (right panel) allows statistical inference based on the deviation from expected flat distribution. Here, the histogram is skewed to the left, which is indicative of the cold bias. It is weak, however, and may be specific to this model-year in accordance with the model's interannual variability. It would not lead to any intervention in operational production. If such a deviation is significant and simultaneously present at many stations, an automated flag would be raised to notify operators of the Climate DT workflow system. When the Climate DT workflow is active at the cadence of one simulated year per day (SYPD), OBSALL performs approximately one on-line monitoring task each calendar-day.

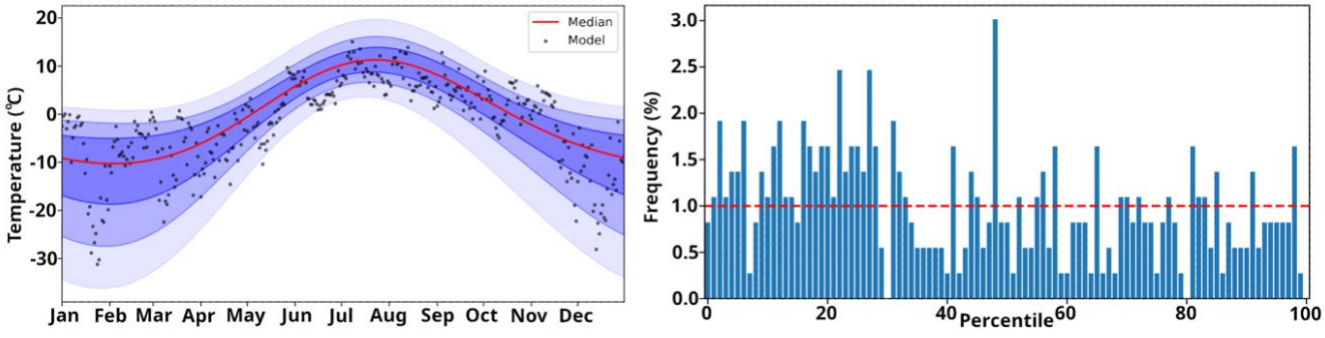


Figure 1. Online monitoring at station 028360 (Sodankylä, Finland). Left panel: simulated values (dots) on the observation-based quantiles (blue shades; shown are 1, 10, 25, 50, 75, 90, and 99 % quantiles); right panel: the same data as a ranked percentile histogram. Simulation: IFS-NEMO historical, 5 km resolution in the atmosphere, year 2014.

4 Posterior evaluation of historic simulations

This Section presents the statistical approach to compute a selection of statistics using the complete ODB as input. We will compare the simulations (Table 1) with synoptic surface observations, as extracted from the Hadley Centre's observational database (HadISD, v3.4.1.202401p; Dunn et al. 2012, 2014, 2016). Free-running model simulations and observations are naturally out-of-sync and therefore a sufficiently long common period is needed for drawing meaningful conclusions. In

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

Climate DT, the posterior evaluation is performed right after a simulation is successfully completed and ODB is fully-augmented with model counterparts in observation-space. The evaluation results in a catalogue of statistics and images that is stored in the repository (Dashboard) for inspection.

4.1 Statistical methods

The observed and simulated climates are characterized by calculating (i) long-term mean monthly, and seasonal and annual (DJF, MAM, JJA, SON, and ANN) means of diurnally averaged values, (ii) average diurnal ranges, (iii) linear trends, (iv) interannual standard deviations of monthly, seasonal, and annual mean values, and (v) standard deviations of irregular variability at sub-monthly time scales. In addition, (vi) extreme value analysis is conducted for yearly extremes and (vii) the occurrence of heatwaves and coldwaves is studied.

All the statistics are calculated from observations and simulation data at 3-hourly intervals (00, 03, …, 21 UTC). Times with missing observations are also omitted from the model simulations, to ensure fair comparison over the common period covering years 1991-2014. In the methods of calculation, as described below, simplicity is in some cases prioritized over statistical preciseness.

1. **Monthly, seasonal, and annual mean values:** Monthly means are first calculated for each 3-hourly UTC time and each year separately. The eight UTC-specific mean values are then averaged to yield the overall monthly mean. If observations for up to four UTC times are missing for the whole month, the overall monthly mean is still calculated, but neglecting the missing UTC times both for the observations and the simulations. Seasonal and annual means are then calculated from the monthly mean values. Finally, the 24-year means of the monthly, seasonal, and annual means are calculated.
2. **Diurnal range**: The diurnal cycle is first defined, for each month and year separately, by the monthly means of each 3-hourly UTC time. The diurnal range is then approximated by the difference between the highest and the lowest of these eight values. This differs in two ways from the most common definition of the diurnal range as the average difference between the daily maximum and minimum. First, the averaging of the UTC-specific values from the beginning to the end of the month reduces the aliasing between the diurnal cycle and irregular weather variability. Second, the eight UTC-specific mean values may not fully cover the range from the minimum to the maximum. For both reasons, this definition yields lower diurnal ranges than the most common definition.
3. **Linear trends.** Monthly, seasonal, and annual means are first computed for each year separately, in the same way as in method 1. The linear trends are then calculated by the ordinary least-squares method from the yearly values.
4. **Interannual standard deviation.** Interannual standard deviation of monthly, seasonal, and annual mean values is calculated as the standard deviation of the regression residuals from method 3.
5. **Intramonth standard deviation**. As a preparation, linear trends from the beginning to the end of each month are calculated separately for each year and each UTC time. Then the standard deviation of the regression residuals is calculated and averaged over the eight UTC times; regression residuals instead of the original values are used to eliminate the

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

contribution from the annual cycle. The seasonal and annual values of this all-UTC standard deviation are obtained by averaging the corresponding monthly values.

6. **Extreme values**. A Gumbel distribution is fit to the annual maxima (for 2-metre temperature, also annual minima) in each year from 1991 to 2014, using the maximum likelihood method. The Gumbel distribution is preferred over the generalized extreme value distribution due to the limited length of the analysis period.
7. **Heatwaves and coldwaves**. A heatwave (coldwave) is defined here as a period of at least three consecutive days with the daily maximum (minimum) 2-metre temperature exceeding (falling below) the 95th (5th) percentile, precomputed for each calendar day using a 31-day centred moving window. By defining events based on percentiles, we ensure that they are identified relative to the specific temperature distribution of observations or models at every location and time of year. This approach makes the metric suitable for comparing extremes across regions with varying climates and among different products. To quantify the intensity of these events, the heatwave magnitude index daily (HWMId; Russo et al. 2015) is employed together with an analogous, novel coldwave magnitude index daily (CWMId) adapted to characterise cold extremes. These indices combine information on both the magnitude and duration of the events, providing an integrated measure of their overall severity rather than relying solely on exceeding the threshold values. Thus, by analysing these events we do not focus solely on peak values, which may be strongly affected by model biases, but also on the persistence of hot and cold episodes across different temperature distributions. The novel coldwave index CWMId is defined as

$$\mathrm{CWMId} = \frac{T_{25p} - T_{min,d}}{T_{75p} - T_{25p}}, \tag{3}$$

where $T_{min,d}$ is the local daily minimum 2-metre temperature of day d and T_{25p} (T_{75p}) is the corresponding 25th (75th) percentile of that calendar day computed from all 31-day centred windows of the 1991–2014 period. For a given coldwave (heatwave), these daily values are later accumulated over the life-cycle of each event to yield CWMIe (HWMIe).

In order to characterize the sampling uncertainty, the observed values in single-station Figures (e.g., in Figure 2) are surrounded by shading (for the annual cycle of monthly values) or error bars (for the seasonal and annual values). In most cases, these cover a range of $\pm 2\sqrt{2}$ standard errors around the observed value. Assuming the same level of variability in the simulations as in the observations, this translates to ±2 standard errors of the model minus observation difference. Thus, when the simulated values are outside the indicated uncertainty range, there is, in broad terms, less than 5% probability that this difference results from chance alone. This method is used for the mean values, diurnal ranges, trends and sub-monthly standard deviations (methods 1, 2, 3, and 5 in the list above). The standard error for the mean values, diurnal range and sub-monthly standard deviation is estimated from the interannual standard deviation of these statistics (s_{obs}) as $\bar{s}_{obs} = s_{obs}/\sqrt{n}$, where n is the number of years and the overbar refers to mean value. For linear trends, the standard error is estimated as

$$s_{trend} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} r_i^2 \Big/ \sum_{i=1}^{n}(year_i - \overline{year_i})^2}\,, \tag{4}$$

where r refers to the regression residual and *year* to the year number.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

For the interannual standard deviation (method 4 in the list), the uncertainty range is inferred from the *F*-distribution. With 22 degrees of freedom for both observations and a model simulation, their standard deviations are significantly different at the 5% risk level when differing by a factor of 1.536. Thus, due to the small sample size, only very large differences in the interannual standard deviation are statistically significant.

For extreme values, a bootstrap procedure is used to estimate the sampling uncertainty. 24 annual extremes are generated by random draws of the cumulative density function of the Gumbel distribution derived from observations. Then the Gumbel distribution is refitted to this random sample and the return values for different return periods are calculated from it. This is repeated 1000 times, and the 2.5…97.5 % range of the resulting return values is shown in the Figures. Note, that this only includes the sampling uncertainty in observations and therefore differs from the other statistics for which the uncertainty ranges also approximately account for the sampling uncertainty in the model simulations.

As a final note, these uncertainty estimates neglect the impact of interannual autocorrelation. This may not always be perfectly justified, but we prefer to accept this source of error in the interest of simplicity.

4.2 Evaluation results

A statistical assessment is presented in this Section on the ability of the Climate DT models to simulate key climate characteristics. The Climate DT models are under continuous development at their host institutions – ECMWF, MPI-M, and AWI – and the evaluation results here provide a snapshot of their current status. Notably, ICON is a relatively new model, whereas the current IFS version benefits from several decades of operational development and interfacing with observations, and thus the two represent totally different development stages. The evaluation is necessarily non-comprehensive, as the primary aim of this paper is to introduce the Climate DT monitoring and evaluation concepts rather than to examine the models' structural choices or relative merits. For that reason, we refrain from drawing comparative conclusions. Among the variables assessed, the simulation of 2-metre temperature is illustrated in somewhat greater detail than the others.

The results are represented as follows. First, detailed analyses, including the annual cycles of observed and simulated climate statistics are presented for three representative stations: station number 028360 Sodankylä in the sub-Arctic Finland (67.395°N, 26.619°N), 081600 Zaragoza in the northern Spain (41.666°N, 1.042°W), and 115180 Prague in the Czech Republic (50.101°N, 14.26°E) (e.g., Fig. 2). These stations are representative of the surrounding areas and far from land-sea contrasts (ensuring nearest-neighbour grid points to have similar physiography in each model). Second, global summary maps of the annual model-minus-observation differences are shown (e.g., Fig. 3). For the extreme value analysis, this approach is slightly modified, so that the return values as a function of return period are shown for the individual stations (e.g., Fig. 8) and the maps give the model-minus-observation differences for 5-year return values (e.g., Fig. S12). Third, the occurrence of heatwaves and coldwaves is reported using cumulative magnitude indices CWMIe and HWMIe, respectively (e.g., Fig. 9).

The six types of statistics were computed for three variables (2-metre temperature and dewpoint difference and 10-metre wind speed). For completeness, the two types of Figures are included for all of them in the Supplementary material. The main body

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

of this article only includes the most central Figures of our study (see Table 2 for orientation). In Table 3, some all-station summary statistics of the model-to-observation differences are given.

Table 2. List of Figures of the mentioned statistics and different variables (T2 = 2-metre temperature; DPD2 = 2-metre dewpoint difference; V10 = 10-metre wind speed); SD = standard deviation.

	Mean values	Diurnal range	Trends	Interannual SD	Intramonth SD	Return values	Heatwaves Coldwaves
T2	2-3 S1-S2	4-5 S3-S4	6 S5-S6	- S7-S8	7 S9-S10	8 S11-S14	9 S15-S16
DPD2	10 S17-S18	- S19-S20	- S21-S22	- S23-S24	- S25-S26	- S27-S28	- -
V10	11 S29-S30	12 S31-S32	- S33-S34	- S35-S36	- S37-S38	- S39-S40	- -

Table 3. Summary statistics from model evaluation across all SYNOP stations. For each variable, statistic, and model, Bias is the average difference between the simulated and observed annual values and MAD the mean absolute difference between the simulated and observed values. %+ gives the percent fraction of stations where the model-minus-observation difference is positive; cases where the difference has the same sign at over 75 % of the stations are highlighted in bold.

		IFS-NEMO			IFS-FESOM			ICON		
		Bias	MAD	%+	Bias	MAD	%+	Bias	MAD	%+
T2 (°C)	Mean	-0.07	0.87	44	-0.15	1.01	46	-0.18	1.89	43
	Diurnal range	-0.47	1.03	32	-0.42	1.06	35	-0.77	1.25	26
	Trend	0.47	0.71	71	0.23	0.72	59	0.30	0.69	66
	Interannual SD	0.05	0.15	59	0.08	0.18	64	0.01	0.14	52
	Intramonth SD	-0.33	0.36	**9**	-0.23	0.32	**20**	-0.29	0.42	**21**
	5-year maximum	-1.16	2.41	37	-0.72	2.46	41	0.75	3.30	55
	5-year minimum	1.10	2.35	68	0.10	2.59	53	-0.72	2.95	46
DPD2 (°C)	Mean	0.82	1.28	59	1.04	1.51	60	1.21	2.18	59
	Diurnal range	0.12	1.11	48	0.18	1.21	46	-1.17	1.47	**17**
	Trend	-0.63	0.98	33	-0.47	0.93	34	-0.55	1.06	34
	Interannual SD	-0.08	0.23	29	-0.08	0.23	29	0.09	0.33	49
	Intramonth SD	-0.12	0.40	31	-0.08	0.43	33	0.18	0.57	55
	5-year maximum	0.18	5.12	46	0.74	5.68	47	2.49	6.47	55
V10	Mean	0.04	0.76	51	0.06	0.78	52	0.23	0.89	56
(m/s)	Diurnal range	-0.77	0.80	**5**	-0.76	0.80	**6**	-0.61	0.73	**14**
	Trend	0.15	0.56	62	0.12	0.57	59	0.10	0.57	59
	Interannual SD	-0.16	0.16	**5**	-0.16	0.16	**5**	-0.14	0.15	**12**
	Intramonth SD	-0.43	0.49	**15**	-0.41	0.49	**17**	-0.06	0.38	41
	5-year maximum	-7.18	7.47	**6**	-6.98	7.27	**6**	-4.09	5.78	**19**

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

		Bias	MAD	%+	Bias	MAD	%+	Bias	MAD	%+

4.2.1 Annual cycle of 2-metre temperature

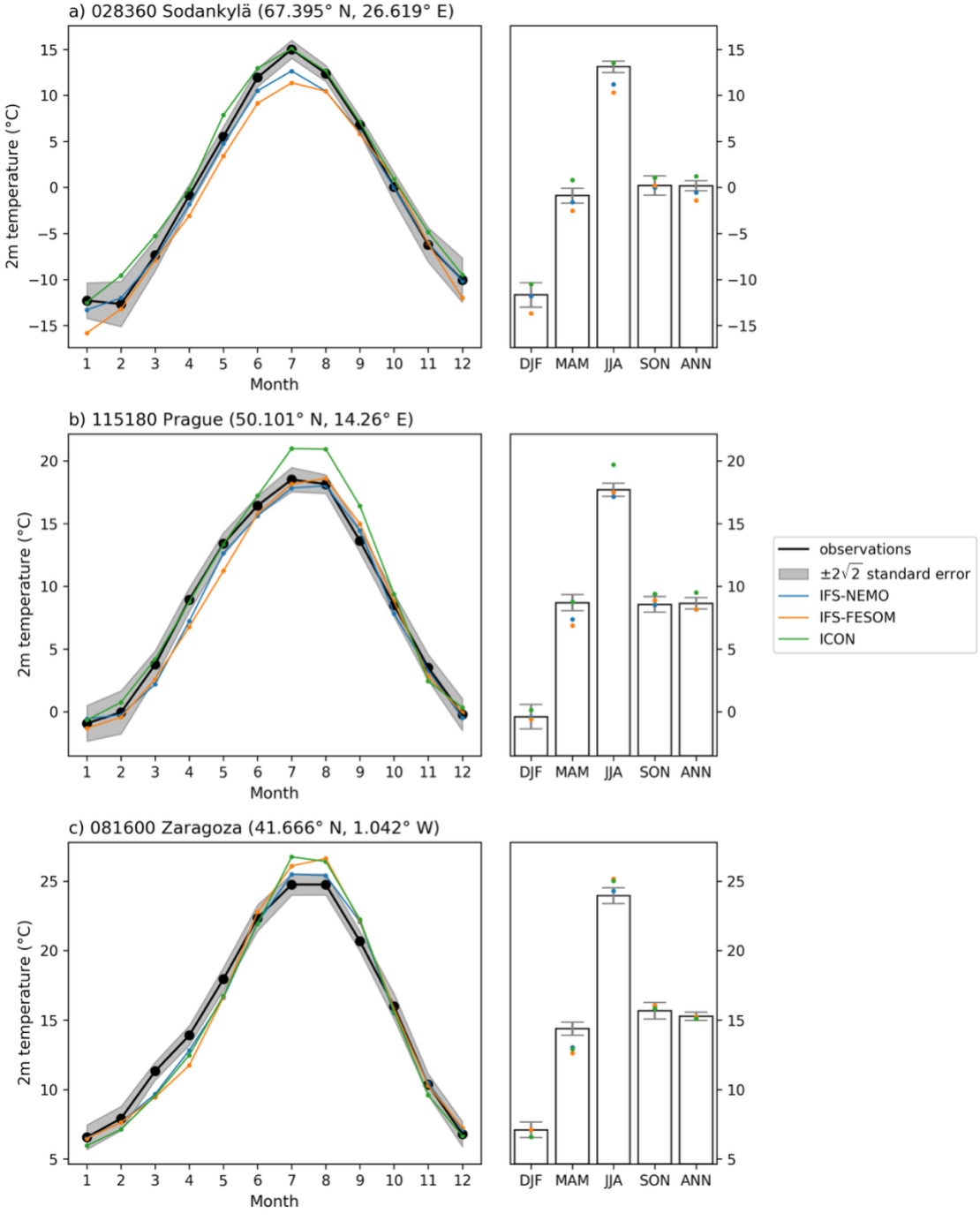
The average annual cycle of temperature is qualitatively well reproduced at every example station (Figure 2). Some biases are evident, including too cold summers in Sodankylä in IFS-FESOM and IFS-NEMO, too warm summers in Prague in ICON, and too cold springs but somewhat too warm summers in Zaragoza in all the simulations. In Prague as well, spring temperatures are significantly below the observed values in IFS-FESOM and IFS-NEMO.

Bias in the annual mean temperature is geographically variable, with a relatively even mix of positive and negative biases in all three simulations (Figure 3). At most stations, the model-to-observation differences appear to be too large to be explained by internal variability alone. The typical magnitude of biases, measured by the mean absolute difference from observations, is nearly twice as large in ICON as in IFS-NEMO and IFS-FESOM (see the map headings in Figure 3 and Table 3). The larger mean absolute difference for ICON reflects both the relatively large negative biases at many tropical stations and the pronounced positive biases in much of extratropical North America and Eurasia.

4.2.2 Diurnal range of 2-metre temperature

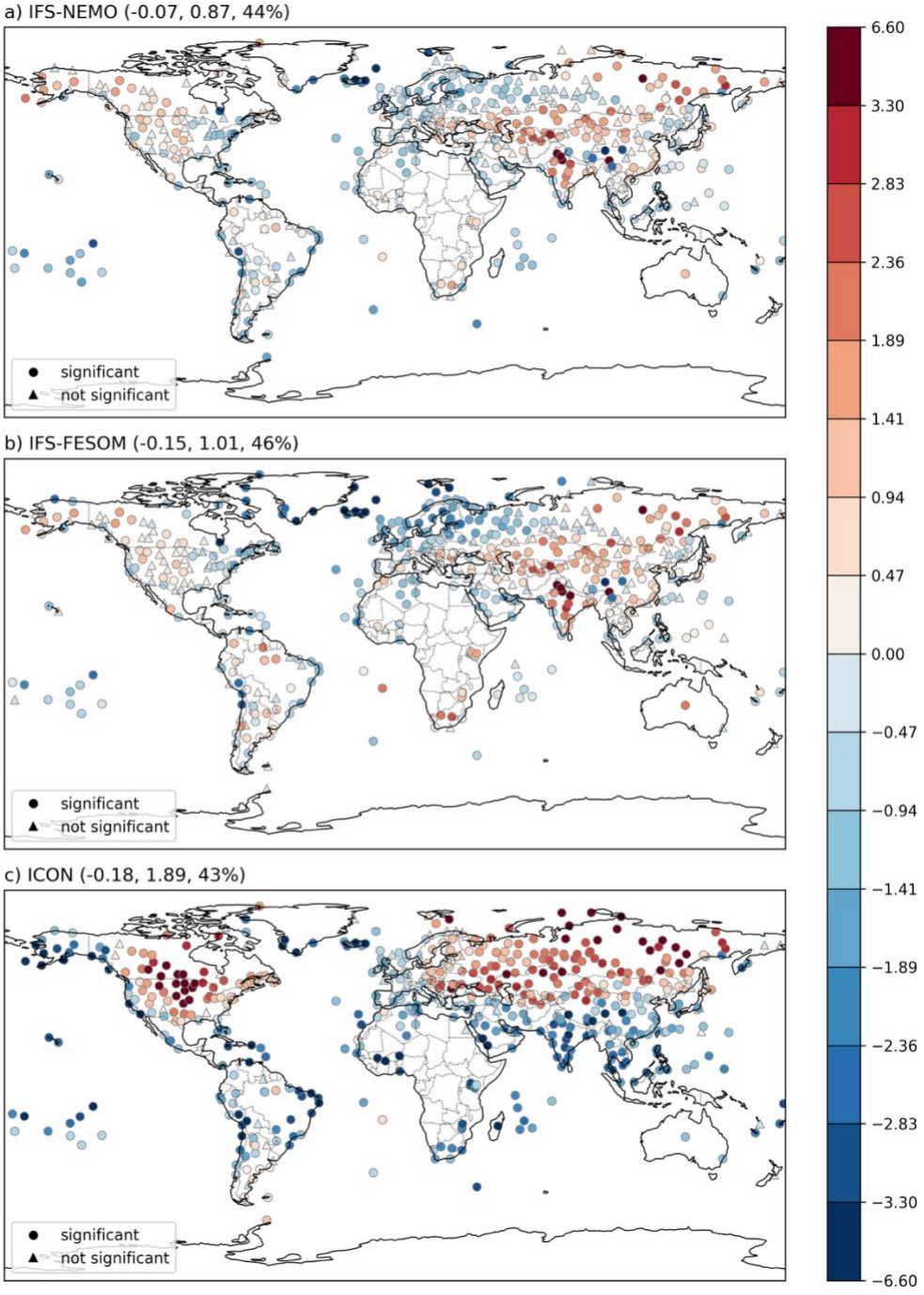
The models correctly capture the near absence of diurnal temperature variability in Sodankylä from November to January, when very little solar radiation is available north of the Arctic circle (Figure 4a). However, the magnitude of the diurnal range is underestimated nearly throughout the year, particularly in ICON. In March, the diurnal range in ICON in Sodankylä is only one third of the observed value, indicating (together with a warm bias in the mean temperature, see Figure 2a) far too mild night temperatures. In Prague, all three models overestimate the diurnal range in autumn and winter (Figure 4b), and the same is also true for IFS-NEMO and IFS-FESOM in Zaragoza (Figure 4c). The biases in spring and summer vary in sign at these stations, even though a clear positive bias in the two IFS models already emerges in summer in Zaragoza. Regarding the global distribution of biases, all three models underestimate the annually averaged diurnal 2-metre temperature range at a majority of the stations (from 65 % in IFS-FESOM to 74 % in ICON; see Figure 5 and Table 3).

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026



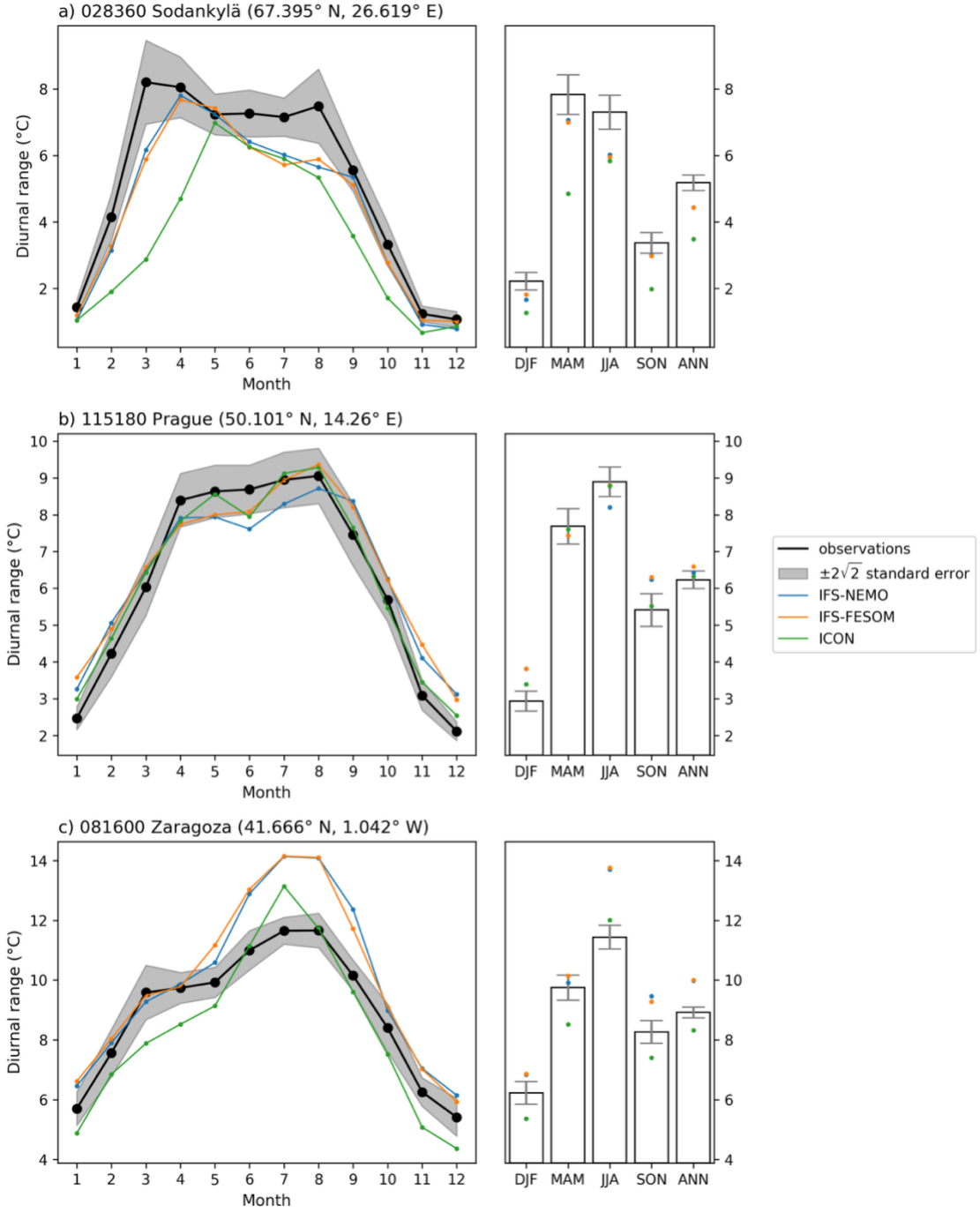
400 **Figure 2**. Left: monthly mean values of 2-metre temperature; stations 028360 Sodankylä (top), 115810 Prague (middle), and 081600 Zaragoza (bottom); observations (black), simulations (coloured lines, see the legend). Right: seasonal and annual mean values in observations (bars) and simulations (coloured circles). Shading on the left and error bars on the right cover $\pm 2\sqrt{2}$ standard errors around the observed value (see Section 4.1).

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026



405 **Figure 3**. Bias of annual mean 2-metre temperature (°C) in (a) IFS-NEMO, (b) IFS-FESOM, and (c) ICON. Stations where the simulated value differ by more (less) than $\pm 2\sqrt{2}$ standard errors from the observations are marked with closed circles (triangles). The first two numeric values in the headings give the average bias and the average absolute bias over all stations, and the last number reports the percent fraction of stations with positive bias.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026



410 **Figure 4**. As Figure 2, but for the diurnal range of 2-metre temperature.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

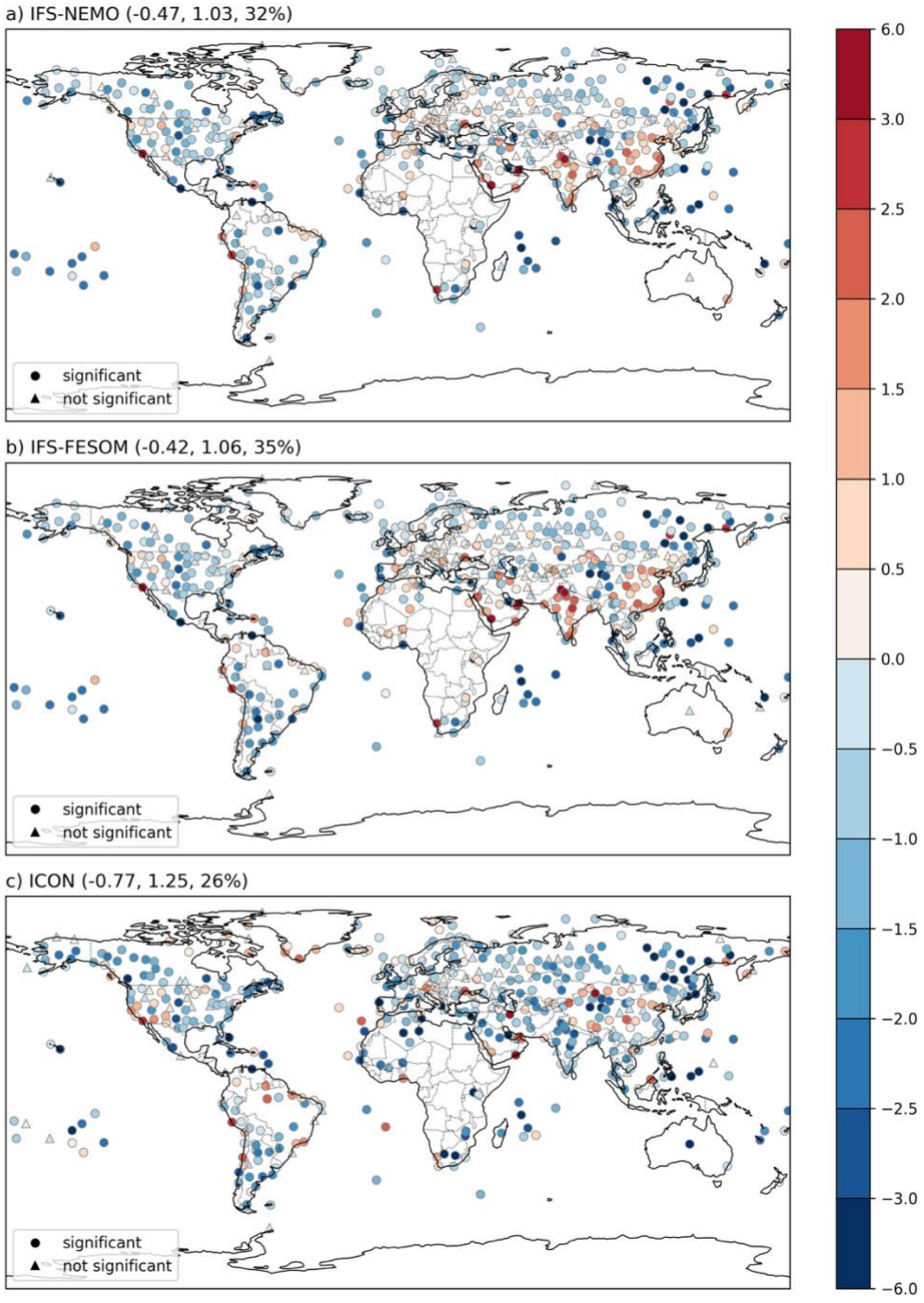


Figure 5. As Figure 3, but for the diurnal range of 2-metre temperature (°C).

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

4.2.3 Trend and interannual variability of 2-metre temperature

All three models overestimate the 1991-to-2014 annual mean warming trend at least at 59–71 % of the individual stations (Figure 6). The bias is largest in IFS-NEMO, in which the all-station mean warming exceeds the corresponding observed value of 0.66°C (23 yr-1) by 0.47°C (23 yr-1) or by about 70 %. Still, the 24-year period is rather short for evaluating the realism of simulated local temperature trends. At extratropical latitudes, in particular, natural variability results in wide error bounds around the observed and simulated trends, and the smaller sample size makes these bounds even wider for monthly-to-seasonal than annual trends (see Figure S5 for an illustration for Sodankylä, Prague and Zaragoza). When and where trend biases of the same sign occur in all three models, this might in some cases just reflect an unusual realization of natural variability in the real world.

The 24-year period is also too short for analysis of interannual variability, meaning that only very large differences between simulated and observed variability can be reliably discerned from noise. For completeness, however, simulated and observed interannual 2-metre temperature variability are compared in Figs. S7-S8.

4.2.4 Intramonth variability of 2-metre temperature

The effective sample size for intramonth variability is much larger than that for interannual variability, allowing more conclusive model evaluation. The intramonth standard deviation of temperature is underestimated in Sodankylä in all the simulations (Figure 7a). However, its annual cycle, with much larger variability in winter than in summer, is correctly simulated. In Prague and Zaragoza as well, the simulated variability in winter and (excluding IFS-FESOM for Zaragoza) spring is smaller than observed. The biases in summer and autumn are less systematic, although IFS-FESOM overestimates the variability in both Prague and Zaragoza in summer.

The tendency of the models to underestimate intramonth temperature variability is widespread. Depending on the model, the simulated annually averaged standard deviation is smaller than observed at 79 % (ICON) to 91 % (IFS-NEMO) of all stations (Table 3 and Figure S10). Exceptions mainly occur on islands and in coastal areas.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

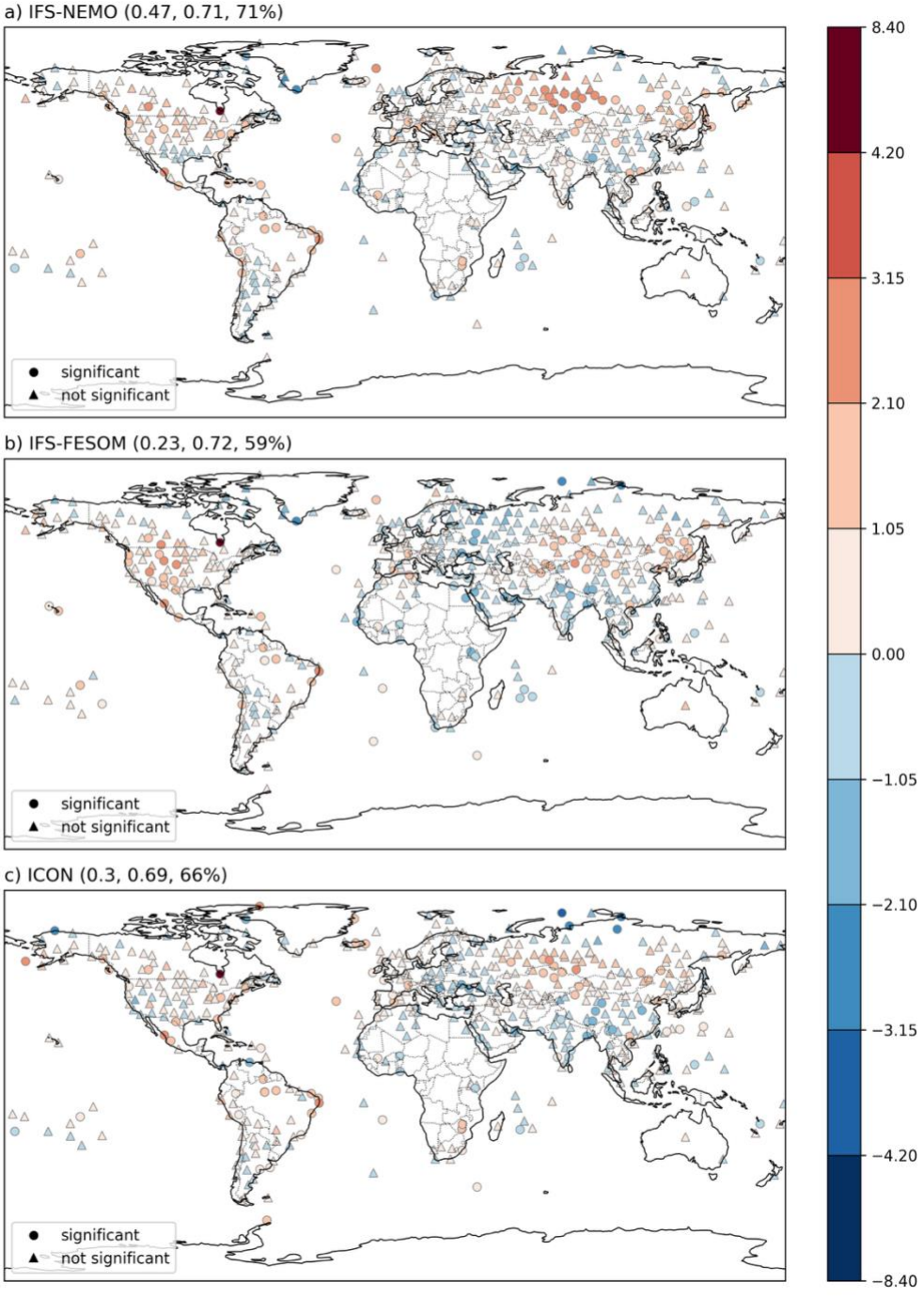
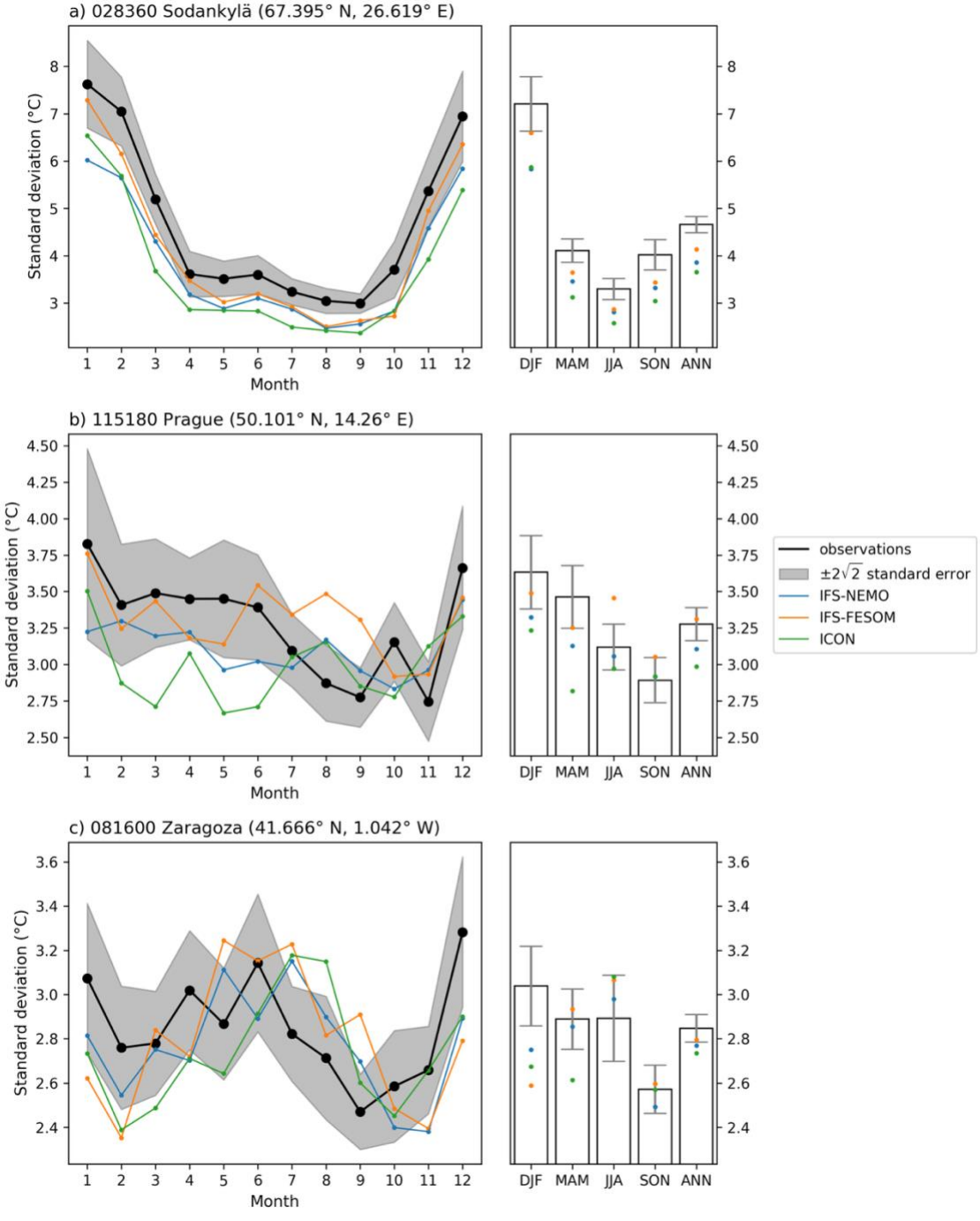


Figure 6. As Figure 3, but for the linear trend of annual mean 2-metre temperature (°C (23 yr^{-1})).

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026



440 **Figure 7**. As Figure 2, but for the intramonth standard deviation of 2-metre temperature.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

4.2.5 Extreme values of 2-metre temperature

The brevity of the 24-year analysis period results in a large sampling variability in extreme values. Nonetheless, as illustrated for Sodankylä in Figure 8, some model-to-observation differences are still apparent. Consistent with the underestimates in both the intramonth variability (Figure 6a), the diurnal range (Figure 4a) and (in IFS-NEMO and IFS-FESOM) the summer mean temperature (Figure 2a), high extremes of temperature at this station are too low in the simulations (Figure 8a). Conversely, the cold extremes are too mild (Figure 8b). Model-minus-observation differences in both the high and the low extremes are geographically variable, but some widespread biases are still apparent (see Figures S12 and S14 for the biases in 5-year return values). For example, while IFS-NEMO and IFS-FESOM underestimate high extremes of temperature at a majority of the stations, ICON tends to overestimate them in the interiors of mid-latitude Eurasia and North America.

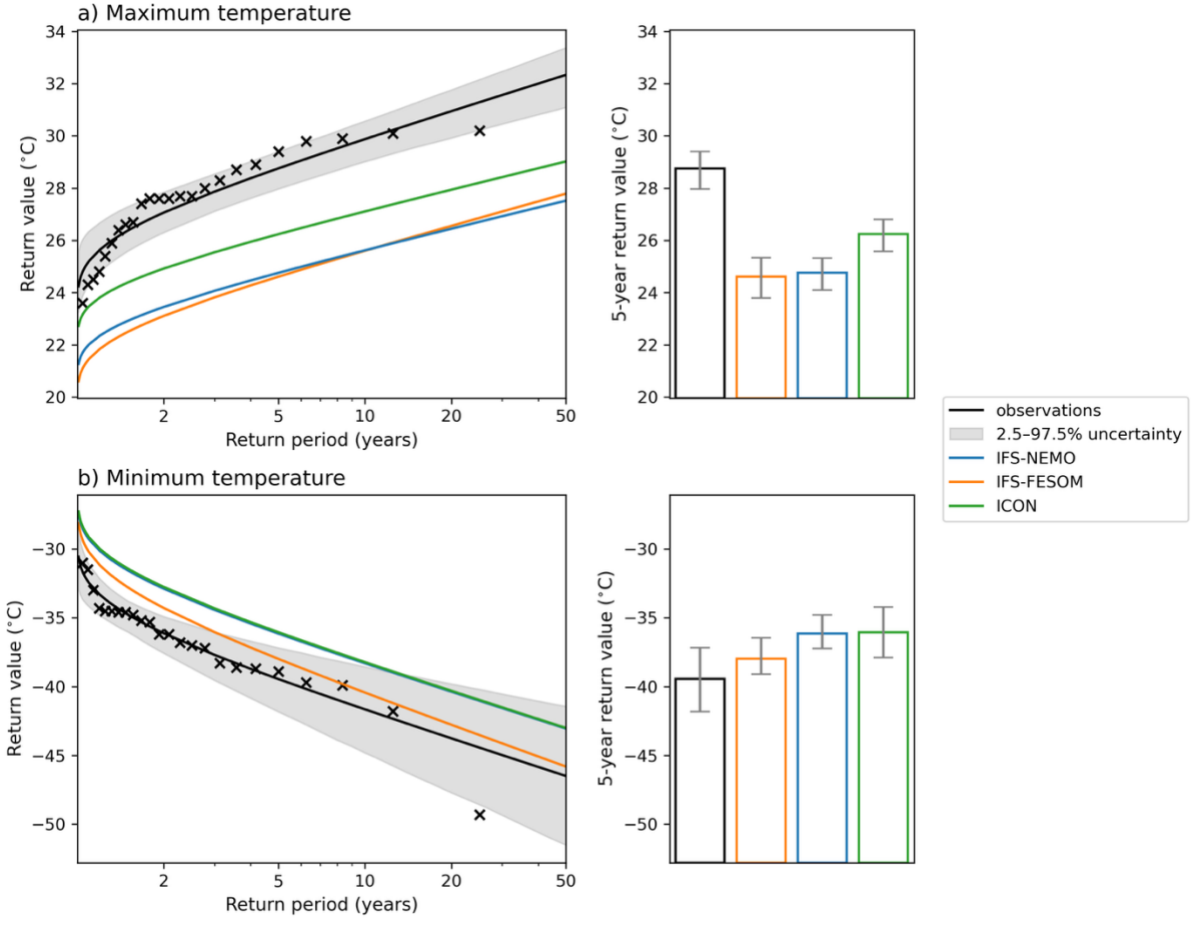


Figure 8. Return values of the highest annual maximum (top) and the lowest minimum 2-metre temperature (bottom) in Sodankylä as a function of the return period, based on a Gumbel distribution fit (black - observations; coloured lines – simulations). The shading shows the 2.5-97.5 % uncertainty range of the observed return values based on 1000 bootstrap samples, and the crosses are the observed annual maximum and minimum values in increasing order of extremity. Right: observed and simulated 5-year return values (bars) and their 2.5-97.5 % uncertainty ranges (error bars).

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

4.2.6 Heatwaves and coldwaves in 2-metre temperature

Figure 9 compares the simulated occurrence of heatwaves and coldwaves with the observational record. In this analysis, we apply quantile-based thresholds to account for differences in temperature distributions between models and observations, reducing the influence of identified biases (such as those in Fig. 2). This approach allows us to focus on evaluating how well the models reproduce persistent heat and cold extremes.

For Sodankylä, the metrics show that the modelled variability generally falls within the observed range, suggesting that the models capture realistic frequencies and intensities of extreme heat and cold. However, closer examination reveals inter-model differences, especially for cold extremes: ICON underestimates the number of events, while IFS-FESOM overestimates them. In terms of temporal patterns, there is no clear trend in heatwave occurrence – only a modest increase in frequency in recent observational years, which the models do not reproduce. By contrast, the number of coldwaves shows a robust decline in both the observations and IFS-NEMO, a trend that is not present in IFS-FESOM or ICON.

To assess the broader applicability of this approach, the analysis was repeated for Prague and Zaragoza (Figures S15 and S16). Despite inter-model variability, a consistent pattern emerges: models generally overestimate the frequency and intensity of extremes – except for IFS-FESOM for heatwaves in Zaragoza. Simulations at both sites suggest more frequent heatwaves and less frequent coldwaves over time, particularly for ICON in Zaragoza (for both types of extremes) and for IFS-NEMO in Prague (for heatwaves). However, these trends are not clearly reflected in the observational records. This suggests that climate simulations may exhibit different variability than observations. Nonetheless, the relatively short 24-year study period limits the ability to draw robust conclusions about extremes.

4.2.7 Annual cycle of 2-metre dew point difference

The 2-metre dew point difference is used here to characterize the humidity of the observed and simulated near-surface climate. In Sodankylä, the annual cycle of the dew point difference is qualitatively well-simulated in IFS-NEMO and IFS-FESOM, although the simulations tend to be slightly too humid in spring and summer while too dry in autumn and winter (Figure 10a). In ICON, however, the average dew point difference is far too low in the winter half-year, falling well below 1°C between December and March. The causes of this clearly unrealistic behaviour would deserve further study.

The dew point differences in Prague and Zaragoza are generally too large, except for spring (Figures 10b-c). This dry bias is particularly pronounced in summer and early fall, and in Prague it is more severe in ICON than in the two IFS models.

Globally, all three models overestimate the dew point difference at about 60 % of the station, with an average bias of the order of 1°C (Table 3 and Fig. S18). In ICON, however, the annual mean dew point difference is too small at 83 % of all stations, with an average bias of -1.14°C.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

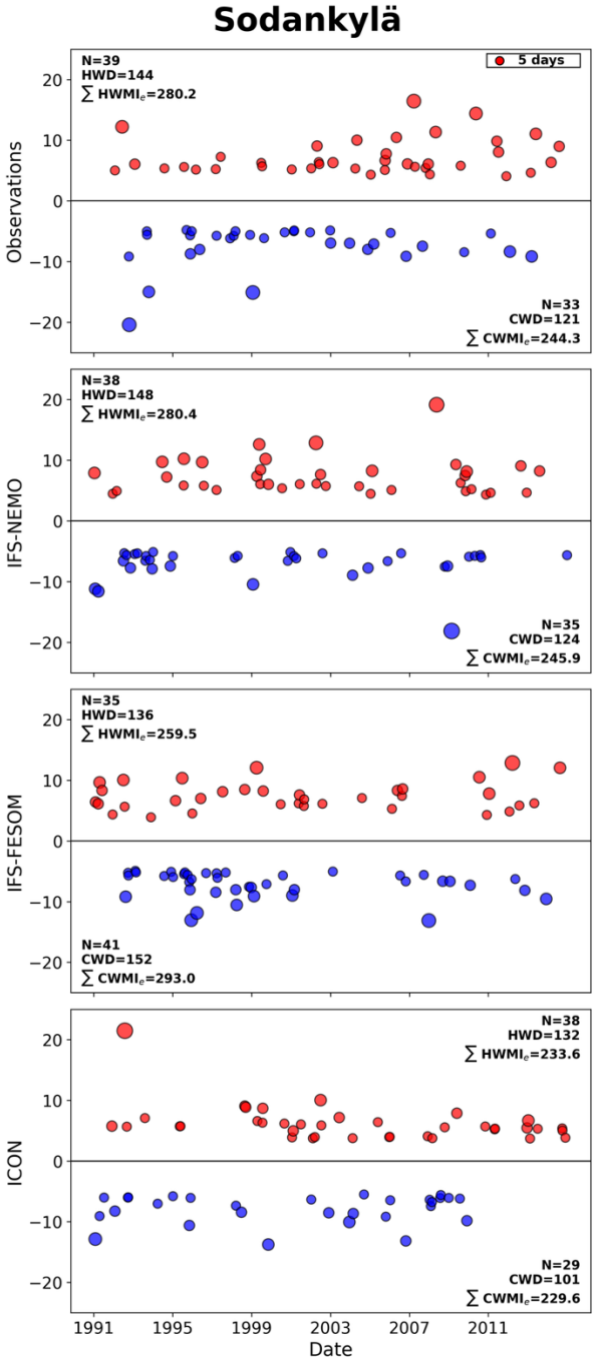


Figure 9. Heatwaves and coldwaves at the station 028360 (Sodankylä, Finland), based on 2-metre temperature observations (top row) and historical simulations from IFS-NEMO (second row), IFS-FESOM (third row), and ICON (fourth row). Red points represent heatwaves ($HWMI_e$), and blue points coldwaves ($CWMI_e$), scaled by –1 for visual symmetry. Point size indicates event duration. In each panel, the top (bottom) portion summarizes heatwave (coldwave) metrics: number of events (N), total number of days (HWD, CWD), and accumulated magnitude indices (Σ $HWMI_e$, Σ $CWMI_e$).

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

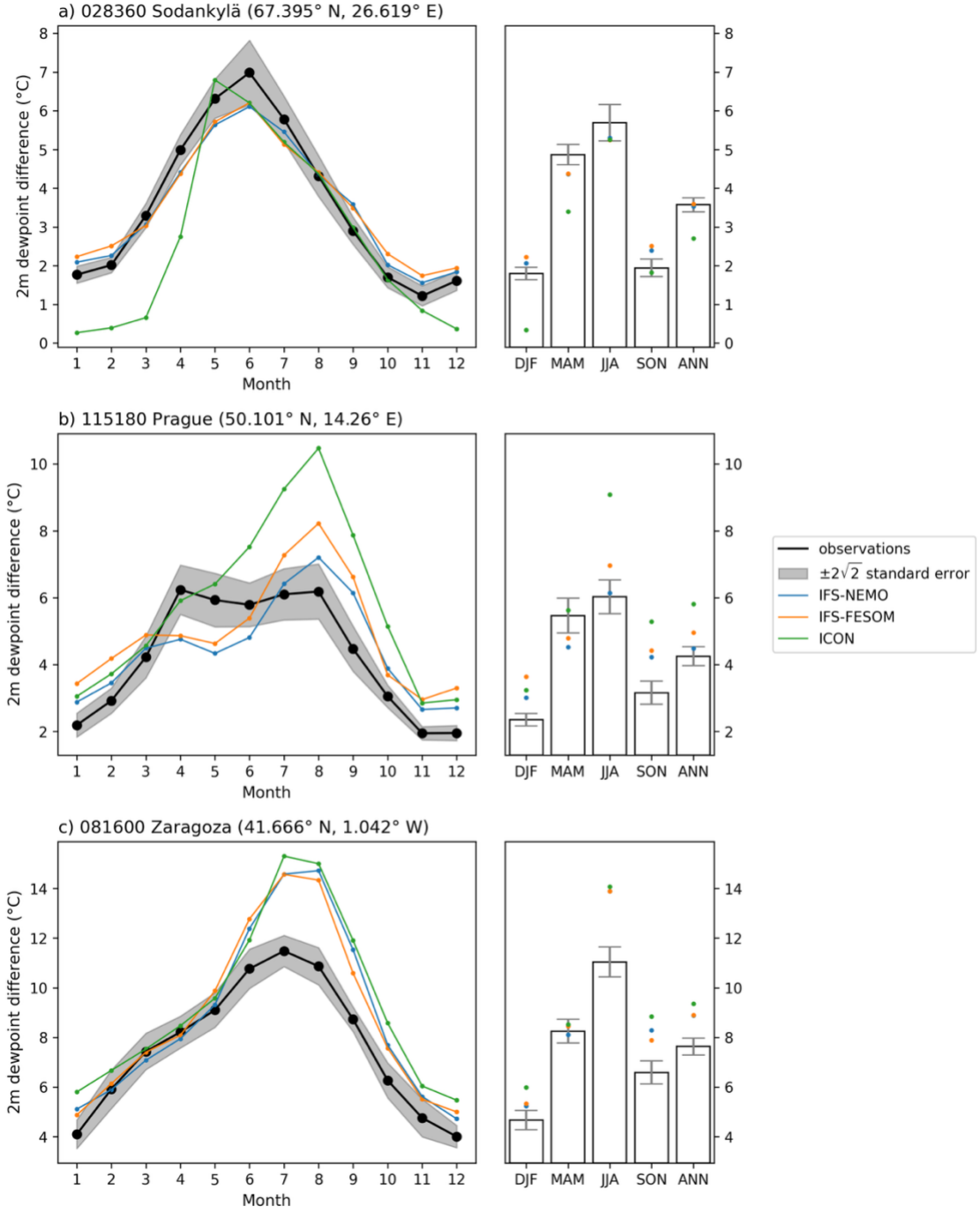


Figure 10. As Figure 2, but for the mean values of 2-metre dew point difference.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

4.2.8 Annual cycle of 10-metre wind speed

The simulation of near-surface wind speed is demanding for any model, due to its dependence on both the atmospheric circulation, boundary layer structure, and local surface characteristics. This is reflected also in Figure 11, which shows a systematic overestimation of wind speed in all the model simulations in Sodankylä (except for summer in ICON), but a clear underestimation in IFS-NEMO and IFS-FESOM in Prague and Zaragoza. In ICON, the annual mean wind speed is closer to the observed values in Prague and Zaragoza, but the difference between winter and summer in Prague is too large. The latter also applies to Sodankylä. Figure S30 shows, among other things, that wind biases are continental in horizontal scales and characterized by negative biases (model wind too weak) at coastal stations in the North and South America.

4.2.9 Diurnal range of 10-metre wind speed

The diurnal range of wind speed is severely underestimated at all three stations, with the exception of Zaragoza in July-August (Figure 12). The bias is pronounced even in Sodankylä, where the simulated mean wind speeds are too strong (Figure 11a). The diurnal range in surface wind speed at inland locations is largely stability-driven, with stronger winds daytime when the boundary layer is less stable and allows stronger turbulent mixing from aloft. Therefore, the underestimation of the diurnal range in 10-metre wind speed points to weaknesses in boundary layer parameterizations that are common to all three models. Similar underestimation extends (with the exception of Sodankylä and ICON in Prague) to the intramonth standard deviation of wind speed (Figure S37) suggesting that, overall, the near-surface winds are not variable enough. The same holds in most cases for extreme wind speeds, with a more severe and ubiquitous underestimation in IFS-NEMO and IFS-FESOM than in ICON (Figure S39).

Globally, a relatively even mix of positive and negative mean wind speed biases occurs in all three models (Table 3 and Figure S30). However, the diurnal range (Figure S32), interannual (Figure S36) and (with the exception of ICON) intramonth standard deviation (Figure S38) as well as the 5-year return value of wind speed (Figure S40) are widely underestimated in all three models.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

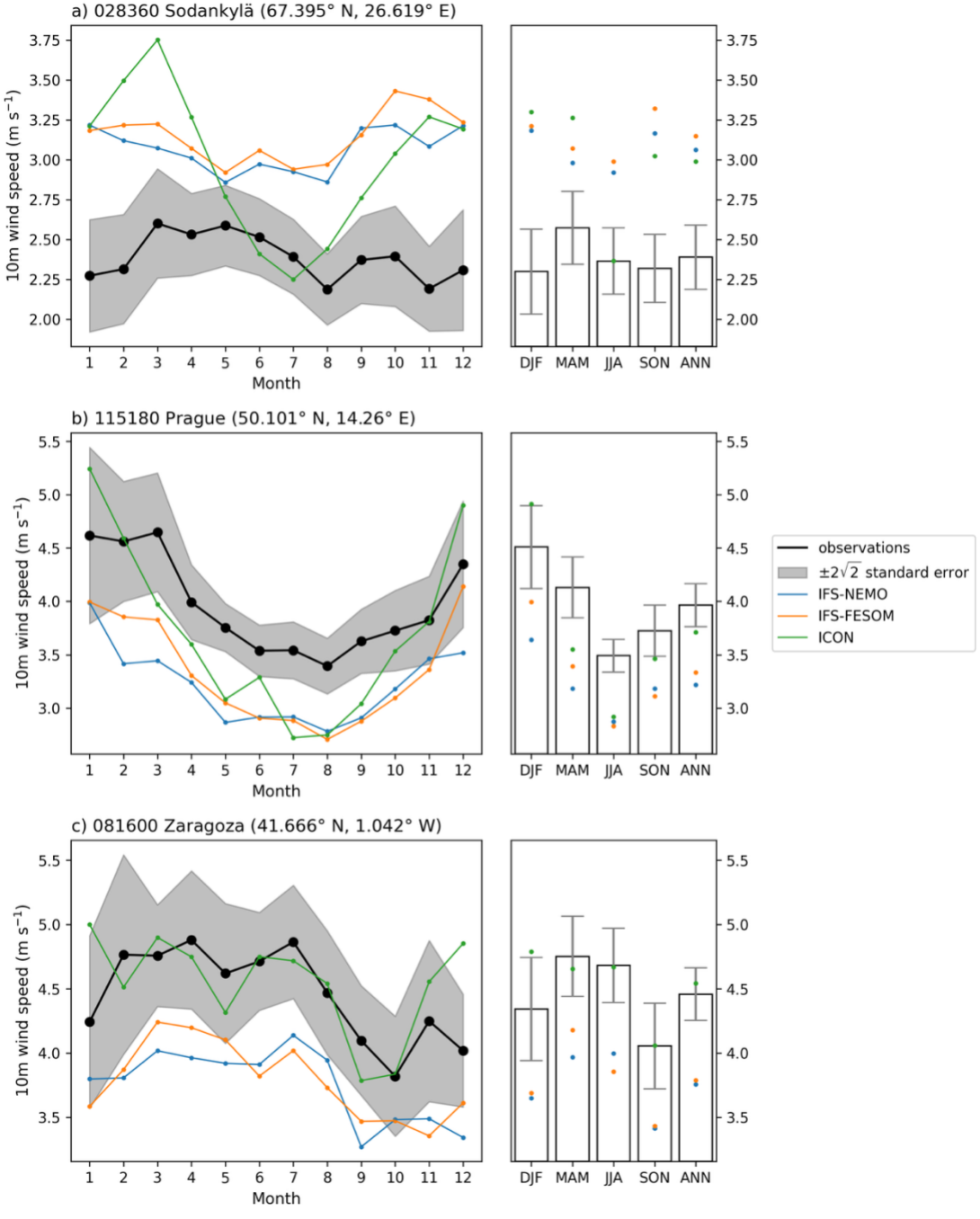


Figure 11. As Figure 2, but for mean values of 10-metre wind speed.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

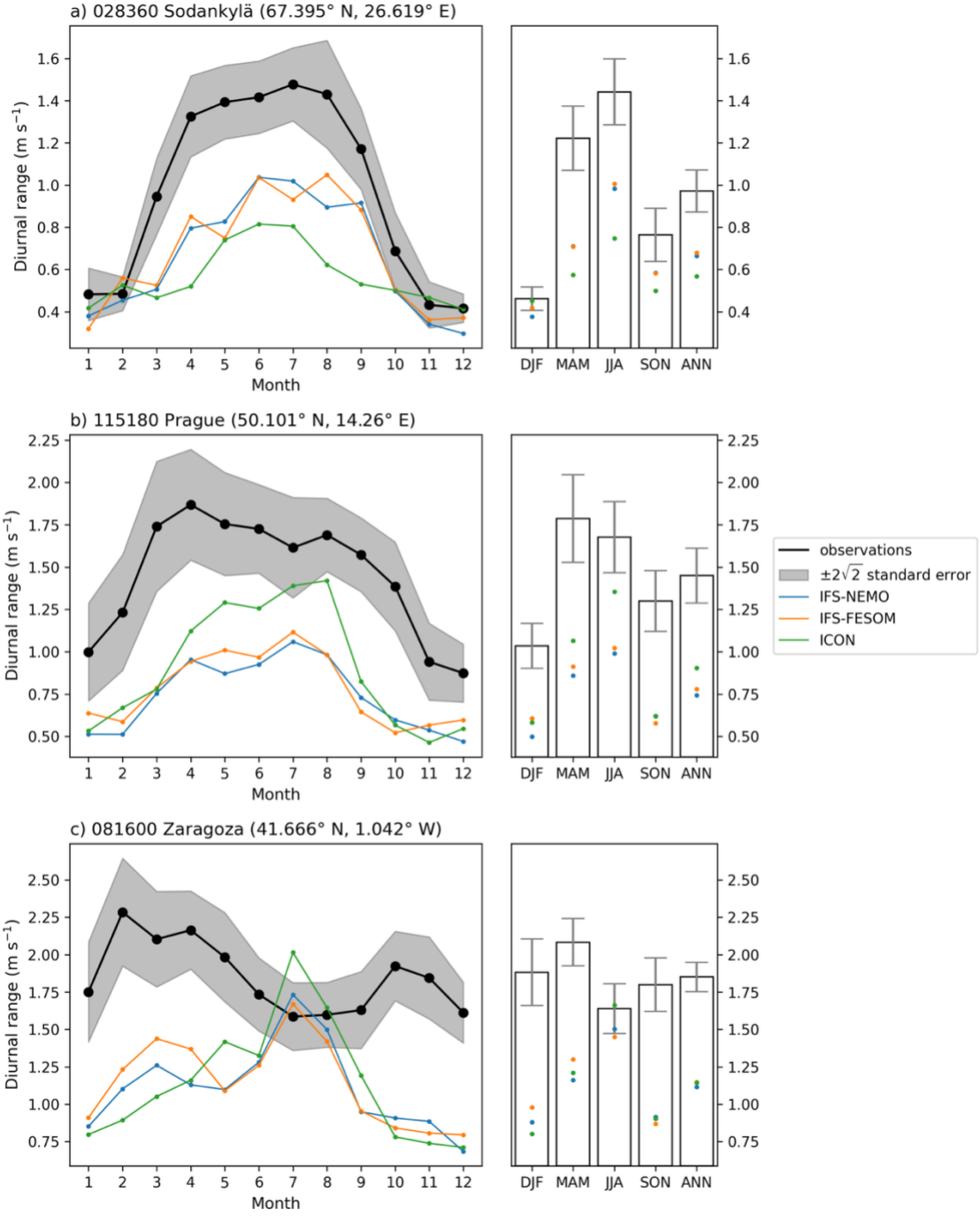


Figure 12. As Figure 2, but for the diurnal range of 10-metre wind speed.

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

5 Discussion

Observation-based evaluation in Climate DT serves two purposes:

1. Run-time (online) monitoring is carried out after each successfully completed model-year, using the corresponding observations of the particular simulation-year. It assesses simulation quality intermittently and aims at the early detection of errors in the simulation setup, rising a flag in case the quality is suspicious. This is reported in Section 3 of this article.
2. Posterior evaluation of simulation quality is performed once the (historic) simulation period is successfully completed. It aims at in-depth statistical analysis and systematic assessment of simulation quality. This is the subject of Section 4.

This two-fold approach offers a key advantage: although the computed observation-space quantities accumulate only gradually in the ODB files as the simulation workflow progresses, even this limited amount of data can already serve as a real-time reality check during execution and, when necessary, alert operators of the Climate DT workflow system. In contrast, the posterior evaluation has access to the complete set of observation-space quantities, enabling more sophisticated data analyses that support the continuous model-development efforts at the host institutes – ECMWF, AWI, and MPI-M – and provide guidance for users of the Climate DT products.

Section 4 presents detailed statistics. The annual cycles of 2-metre temperature – and, to a somewhat lesser extent, humidity and 10-metre wind speed – are generally well captured by all models, exhibiting only minor biases in their monthly mean values. This successful representation of key climate variables indicates a well-balanced simulation of dynamical processes across the coupled atmosphere–ocean system. A contributing factor to the small overall biases is the tuning process, in which dominant climate drivers are harmonised; for example, models' free parameters are adjusted to achieve a realistic top-of-atmosphere radiative balance and energetic consistency while maintaining minimal climate drift (Mauritzen et al. 2012, 2022; Bechtold et al. 2000). Climate DT, however, employs computationally expensive modelling configurations, leaving limited scope for dedicated fine-tuning.

Overall, the Climate DT models tend to under-represent process-level variability. This is particularly pronounced for 10-metre wind speed across all examined time-scales (diurnal, intramonth, and interannual), and for 2-metre temperature at intramonth and – more weakly – diurnal scales. However, this systematic under-representation does not extend to 2-metre temperature at interannual scales (Figs. S7 and S8), to heat/coldwaves (Figs. 9, S15, and S16), or to 2-metre dew-point difference at any time-scale (Figs. S20, S24, S26). The absence of certain aspects of variability in climate models is well documented (Wang and Clow 2020; Zha et al. 2023; Zhan and Wang 2024; Halifa-Marin et al. 2025). The novelty here is the demonstration that this behaviour also appears in storm-resolving climate models with kilometre-scale processes, models that can reasonably be expected to capture local climate variability.

Some of the statistics, particularly the diurnal range and irregular variability of surface temperature, humidity, and wind, appear to be sensitive to process-level modelling uncertainties. As a result, systematic errors tend to accumulate in these metrics, indicating sub-optimal representation of the planetary boundary layer, soil hydrology and thermodynamics, and their coupling processes. It is also plausible that synoptic-scale variability in the free atmosphere does not propagate effectively into surface

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

variables, thereby reducing irregular variability on sub-monthly time-scales, even when mid-tropospheric variability is well captured.

ODB files naturally enable broader scientific exploration of the Earth's physical climate system. While this paper focuses on the observed historical period, the observation projections can, in principle, be extended into the unobserved future period of the simulations by applying an assumed observational network – for example, by re-using station locations and observed variables from the historical era. Such extensions support investigations into how the climate-change signal manifests within observation space, including identifying the point at which a simulation departs from the envelope defined by observed quantiles (i.e., the time of emergence of climate change; Hawkins and Sutton 2012).

One limitation of OBSALL v1.0 is the restricted set of observing systems it includes. Ideally, the full suite of Earth observations used in global reanalysis systems such as ERA5 would provide an almost unconstrained foundation for climate research. Such a diverse set of Earth observations could support a comprehensive effort to diagnose and reduce systematic errors in free-running models. Beyond error identification, such an expanded ODB framework would also enable observation-driven algorithmic model tuning, thereby reducing parametric uncertainties in future model generations (Ekblom et al. 2023).

Both the ODB archive and the broader observation-projection framework are designed to be readily extendable. ODB is an extremely space-efficient and fast storage system, and its use within OBSALL leverages synergies with the advanced data-handling infrastructures at ECMWF. Because kilometre-scale simulations are computationally expensive and generate vast volumes of output, the observation projection becomes the only long-term archive retaining information at native storm-resolving resolution. In this sense, ODB can be viewed as a dimension-reduced representation of the full-resolution simulation data. This compact yet information-rich format has the potential to serve specific user groups effectively, including those working with climate adaptation at local (city) scales.

The Climate DT provides an operational simulation framework that supports climate-change adaptation, with observation-space computations playing a central role. First, the evaluation of Climate DT models serves as observation-based uncertainty quantification, addressing the reliability of the information delivered to users. Second, it enables the development of user products expressed directly in observation space – an intuitive, down-to-earth way of communicating adaptation-relevant information. Looking ahead, these capabilities naturally extend to storyline simulations within the emerging near-real-time framework of Climate DT, where observation-based assessment will likewise remain essential.

6 Conclusions

Destination Earth Climate Change Adaptation Digital Twin (Climate DT) has advanced systems for monitoring simulation quality: AQUA (Nurisso et al. 2026), using gridded reference data, and OBSALL, using raw Earth observations. The observation-based monitoring system is implemented into the Climate DT workflow system, which allows to evaluate simulation quality "on-the-fly" using the run-time access to the full-resolution simulation data. By applying state-of-the-art

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

observation operators, adopted from numerical weather prediction, OBSALL informs the Climate DT simulation system with potentially any available Earth observing system.

This article focuses on the theoretical foundation underpinning observation-based online monitoring and posterior model evaluation. The technical implementation is also explained together with the evaluation results of the Climate DT Phase 2 simulations, as seen through the lens of high-quality synoptic surface observations.

The evaluation covers the simulated mean, trend, variability and extremes of historic simulations from 1991 to 2014 of the IFS-NEMO, IFS-FESOM, and ICON models. The main findings are as follows. The annual cycles of 2-metre temperature and (to a somewhat lesser extent) humidity and 10-metre wind speed are generally well captured by all models, with minor biases in their monthly mean values. More refined statistics, such as the diurnal range of surface temperature, humidity, and wind speed – statistics that are sensitive to process-level modelling uncertainties – are indicative of systematic model errors, which are plausibly related to modelling of boundary layer and soil processes, and their mutual couplings. As a general conclusion, the observation-space variables tend to have too little process level variability at diurnal, intramonth, and interannual time-scales, especially regarding 10-metre wind speed but also in 2-metre temperature.

Presentation of the OBSALL concept underscores the usefulness of online observation projection in both run-time monitoring and posterior evaluation of climate simulations. The paper illustrates how Earth observations can be used to inform on digital twin technologies. It is a novelty that Climate DT simulation models can now be examined at such level of details and objectively interfaced with raw Earth observations containing the corresponding process-level imprints. The evaluation results indicate that while there is still work to be done to improve the simulation models, the continuous model development process has gained a valuable new resource. Furthermore, since climate change adaptation mostly occurs at local level, the raw observation-based evaluation informs directly the scales of interest. Climate DT simulation data archive will contain a progressively increasing number of experiment-specific observation database files, which allow convenient posterior evaluation and inter-comparison over generations of models, thus improving understanding how adaptation information evolves regarding local level decision-making.

In summary, the Destination Earth Climate Change Adaptation Digital Twin (https://platform.destine.eu/) takes a significant step to integrate Earth observing systems and observation modelling capabilities into operational climate simulation workflow. The novel solution for projecting climate simulations into observation space is a step beyond the state-of-the-art in observation-based climate model evaluation, thanks to its highly progressive integrated workflow.

Code and data availability

Except where specifically noted, the software used is open source under the licences listed in the relevant software archive. The Climate DT dataset is accessible via https://doi.org/10.21957/d3f982672e (DestinE, 2025). There is a unique semantic data access for the data that allows users to clearly delineate the contributing models, experiment, dates, and variables, among other parameters. The availability policy of datasets in the data lake is defined in https://destine-data-lake-

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

docs.data.destination-earth.eu/en/latest/dedl-discovery-and-data-access/DestinE-Data-Policy-for-DestinE-Digital-Twin-Outputs/DestinE-Data-Policy-for-DestinE-Digital-Twin-Outputs.html (last access: 20 May 2026). Access requires registration to the platform https://platform.destine.eu and applying for upgraded access. Details are provided in the FAQs about the data access https://platform.destine.eu/support-pages/data-access. Autosubmit is available at https://github.com/BSC-ES/autosubmit (last access: 20 May 2026) and the version used in the manuscript is available at https://doi.org/10.5281/zenodo.15590529 (Beltrán Mora et al., 2026; Manubens-Gil et al., 2016). The Climate DT workflow is available in https://github.com/DestinE-Climate-DT/Workflow (last access: 20 May 2026), and the version used is archived available at https://doi.org/10.5281/zenodo.15607598 (Arriola et al., 2025; Manubens-Gil et al., 2016). OBSALL is available at https://doi.org/10.5281/zenodo.15628903 (Tuppi et al., 2025). The version used here of the OBSALL python code with full documentation is available under Apache License 2.0 on GitLab at https://earth.bsc.es/gitlab/digital_twins_public/obsall. Scripts and data to reproduce the Figures of this manuscript can be found at [to-be-specified].

Author contributions

HJ conceptualized the approach and led the implementation with substantial support from all coauthors; HJ and JR wrote the original draft; LT developed the technical solution for computing the observation-space quantities (observation projection) and run-time monitoring; JR and AS-B developed the statistical methodology for posterior evaluation; CB developed the OBSALL workflow solution; JT and AT contributed to the data analysis and visualisation. PD, FD-R, TJ, DK, JK, SB, and IS have commented and reviewed progress versus plans of OBSALL development in Climate DT Phase 1 and 2. All authors contributed to the manuscript in discussions, and read, commented, and approved the final manuscript.

Competing interests

The corresponding author declares that none of the authors has any competing interests.

Disclaimer

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Acknowledgements

The work presented in this paper has been produced in the context of the European Union's Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts implementing part of this Initiative with funding by European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. We acknowledge the EuroHPC Joint Undertaking (JU) for awarding

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

this project access to the EuroHPC supercomputer LUMI and MareNostrum5 through a EuroHPC JU Special Access call. AS-B was supported by the Helmholtz Research Field Earth & Environment for the Innovation Pool Project ACTUATE.

References

Andersson E and H Järvinen (1999) Variational quality control. Quarterly Journal of the Royal Meteorological Society, 125, 697-722. https://doi.org/10.1002/qj.49712555416

Arriola, L, Gaya i Àvila, A, Roura Adserias, F, et al. (2025) DestinE-Climate-DT/Workflow: v5.1.2 (v5.1.2), Zenodo [code], https://doi.org/10.5281/zenodo.15607598

Bauer, P, Stevens, B and W Hazeleger (2021) A digital twin of Earth for the green transition. Nature Climate Change, 11, 80–83. https://doi.org/10.1038/s41558-021-00986-y

Beltrán Mora, D, Kinoshita, B P, Marciani, M G, et al. (2026) Autosubmit (v4.1.14), Zenodo [code], https://doi.org/10.5281/zenodo.15590529

Bock, L et al. (2020) Quantifying Progress Across Different CMIP Phases with the ESMValTool. Journal of Geophysical Research: Atmospheres, 125, e2019JD032321. https://doi.org/10.1029/2019JD032321

Dee, D, Obregon, A and C Buontempo (2024) Are Our Climate Data Fit for Your Purpose? Bulletin of the American Meteorological Society, 105, E1723–E1733. https://doi.org/10.1175/BAMS-D-23-0295.1

Doblas-Reyes, F J, Kontkanen, J, Sandu, I, et al. (2026) The Destination Earth digital twin for climate change adaptation. Geoscientific Model Development, 19, 2821–2848. https://doi.org/10.5194/gmd-19-2821-2026

Dunn, R J H, Willett, K M, Thorne, P W, et al. (2012) HadISD: a quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. Climate of the Past, 8, 1649-1679. https://doi.org/10.5194/cp-8-1649-2012

Dunn, R J H, Willett, K M, Morice, C P and D E Parker (2014) Pairwise homogeneity assessment of HadISD. Climate of the Past, 10, 1501-1522. https://doi.org/10.5194/cp-10-1501-2014

Dunn, R J H, Willett, K M, Parker, D E and L Mitchell (2016) Expanding HadISD: quality-controlled, sub-daily station data from 1931. Geoscientific Instrumentation, Methods and Data Systems, 5, 473-491. https://doi.org/10.5194/gi-5-473-2016

Durre, I et al. (2006) Overview of the Integrated Global Radiosonde Archive. Journal of Climate, 19, 53-68. https://doi.org/10.1175/JCLI3594.1

Ekblom, M, Tuppi, L, Räty, O, Ollinaho, P, Laine, M and H Järvinen (2023) Filter Likelihood as an Observation-Based Verification Metric in Ensemble Forecasting. Tellus A: Dynamic Meteorology and Oceanography, 75(1): 69–87. https://doi.org/10.16993/tellusa.96

Eyring, V, Bony, S, Meehl, G A, et al. (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9, 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

Eyring, V et al. (2020) Earth System Model Evaluation Tool (ESMValTool) v2.0-an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. Geoscientific Model Development, 13, 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Frick, C, Steiner, H, Mazurkiewicz, A, et al. (2014) Central European high-resolution gridded daily data sets (HYRAS): Mean temperature and relative humidity. Meteorologische Zeitschrift, 23, 15–32. doi:10.1127/0941-2948/2014/0560

Fouilloux, A (2009) ODB (Observational DataBase) and its usage at ECMWF. Twelfth Workshop on Meteorological Operational Systems, 2-6 November 2009. 5 pp. https://www.ecmwf.int/en/elibrary/74516-odb-observational-database-and-its-usage-ecmwf

Gampe, D, Schmid, J and R Ludwig (2019) Impact of Reference Dataset Selection on RCM Evaluation, Bias Correction, and Resulting Climate Change Signals of Precipitation. Journal of Hydrometeorology, 20, 1813–1828. https://doi.org/10.1175/JHM-D-18-0108.1

Grayson, K, Thober, S, Lacima-Nadolnik, A, et al. (2025) Statistical summaries for streamed data from climate simulations: one-pass algorithms. Geoscientific Model Development, 18, 5873–5890. https://doi.org/10.5194/gmd-18-5873-2025

Gorski, K M, Hivon, E, Banday, A J, et al. (2005) HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. The Astrophysical Journal, 622, 759–771. https://doi.org/10.1086/427976

Halifa-Marín, A, Torres-Vázquez, M A, Trigo, R, Vicente-Serrano, S M, Turco, M, Jiménez-Guerrero, P, et al. (2025) Too-stable North Atlantic climate system in CMIP6 experiments undermines precipitation projections in Europe. Quarterly Journal of the Royal Meteorological Society, 151, e4999. https://doi.org/10.1002/qj.4999

Hawkins, E and R Sutton (2012) Time of emergence of climate signals. Geophysical Research Letters, 39, L01702. https://doi.org/10.1029/2011GL050087

Hersbach, H, Bell, B, Berrisford, P, et al. (2020) The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049. https://doi.org/10.1002/qj.3803

Hodson, T (2025) Python interface for ODC. https://github.com/ecmwf/pyodc

Hohenegger, C, Korn, P, Linardakis, L, et al. (2023) ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. Geoscientific Model Development, 16, 779–811. https://doi.org/10.5194/gmd-16-779-2023

Jakob, C, Gettelman, A and A Pitman (2023) The need to operationalize climate modelling. Nature Climate Change, 13, 1158–1160. https://doi.org/10.1038/s41558-023-01849-4

John, A, Beyer, S, Athanase, M, Sánchez-Benítez, A, Goessling, H F, Hossain, A, et al. (2026) Global kilometer-scale climate storylines using spectral nudging. Journal of Advances in Modeling Earth Systems, 18, e2025MS005326. https://doi.org/10.1029/2025MS005326

Kaipio, J P and E Somersalo (2005) Statistical and Computational Inverse Problems. Applied Mathematical Sciences, Volume 160. Springer New York, NY. ISBN978-0-387-22073-4. https://doi.org/10.1007/b138659

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

Kalnay, E (2002) Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press. ISBN 9780511802270. https://doi.org/10.1017/CBO9780511802270

Lean, P, Holm, E V, Bonavita, M, et al. (2021) Continuous data assimilation for global numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 147, 273-288. https://doi.org/10.1002/qj.3917

Leuridan, M, Hawkes, J, Smart, S, et al. (2025) Polytope: an algorithm for efficient feature extraction on hypercubes. Journal of Big Data, 12, 243. https://doi.org/10.1186/s40537-025-01306-3

Manubens-Gil, D, Vegas-Regidor, J, Prodhomme, C, Mula-Valls, O and J F Doblas-Reyes (2016) Seamless management of ensemble climate prediction experiments on HPC platforms. In: 2016 International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, Austria, 895–900. https://doi.org/10.1109/HPCSim.2016.7568429

Moreno-Chamarro, E, Caron, L-P, Loosveldt Tomas, S, et al (2022) Impact of increased resolution on long-standing biases in HighResMIP-PRIMAVERA climate models. Geoscientific Model Development, 15, 269–289. https://doi.org/10.5194/gmd-15-269-2022

Nurisso, M, Caprioli, S, Davini, P, et al. (2025) AQUA. https://doi.org/10.5281/zenodo.15044749

Nurisso, M, von Hardenberg, J, Cadau, M, Caprioli, S, Ghinassi, P, Ghosh, S, Koldunov, N, Nazarova, N, Rajput, M M, Tovazzi, E and P Davini (2026) AQUA v1: The Application for QUality Assessment for the Climate Change Adaptation Digital Twin, EGUsphere [preprint]. https://doi.org/10.5194/egusphere-2026-1115

Pailleux, J (1990) A global variational assimilation scheme and its application for using TOVS radiances. In: Proceedings of the WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography, pp. 325–328, Clermont-Ferrand, France.

Rabier F, Järvinen H, Klinker E, et al. (2000) The ECMWF implementation of four dimensional variational assimilation. Part I: Experimental results with simplified physics. Quarterly Journal of the Royal Meteorological Society, 126, 1143-1170. https://doi.org/10.1002/qj.49712656415

Rackow, T, Pedruzo-Bagazgoitia, X, Becker, T, et al. (2025) Multi-year simulations at kilometre scale with the Integrated Forecasting System coupled to FESOM2.5 and NEMOv3.4. Geoscientific Model Development, 18, 33–69. https://doi.org/10.5194/gmd-18-33-2025

Raoult B (1997) Architecture of the new MARS server. In: Sixth Workshop on Meteorological Operational Systems, 17-21 November 1997. https://www.ecmwf.int/en/elibrary/76107-architecture-new-mars-server

Russo, S, Sillmann, J and E M Fischer (2015) Top ten European heatwaves since 1950 and their occurrence in the coming decades. Environmental Research Letters, 10, 124003. https://doi.org/10.1088/1748-9326/10/12/124003

Sánchez-Benítez, A, Goessling, H, Pithan, F, et al. (2022) The July 2019 European Heat Wave in a Warmer Climate: Storyline Scenarios with a Coupled Model Using Spectral Nudging. Journal of Climate, 35, 2373–2390. https://doi.org/10.1175/JCLI-D-21-0573.1

Sandu, I (2024) Destination Earth's digital twins and Digital Twin Engine – state of play. ECMWF Newsletter, 14–23. https://doi.org/10.21957/is1fc736jx

https://doi.org/10.5194/egusphere-2026-3364
Preprint. Discussion started: 24 June 2026

Segura, H, Pedruzo-Bagazgoitia, X, Weiss, P, et al. (2025) nextGEMS: entering the era of kilometer-scale Earth system modeling, EGUsphere, 2025, 1–39. https://doi.org/10.5194/egusphere-2025-509

Shen, C et al. (2022) Evaluation of global terrestrial near-surface wind speed simulated by CMIP6 models and their future projections. Annals of the New York Academy of Sciences, 1518, 249-263. https://doi.org/10.1111/nyas.14910

Shepherd, T G, Boyd, E, Calel, R A, et al. (2018) Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. Climatic Change, 151, 555–571. https://doi.org/10.1007/s10584-018-2317-9

Stevens, B (2024) A Perspective on the Future of CMIP. AGU Advances, 5, e2023AV001086. https://doi.org/10.1029/2023AV001086

Stocker, E F, Alquaied, F, Bilanow, S, et al. (2018) TRMM Version 8 Reprocessing Improvements and Incorporation into the GPM Data Suite. Journal of Atmospheric and Oceanic Technology, 35, 1181–1199. https://doi.org/10.1175/JTECH-D-17-0166.1

Tuppi, L, Bouvier, C, Räisänen, J and H Järvinen (2025) Observation operators for climate models (OBSALL), in: The Destination Earth digital twin for climate change adaptation, Zenodo [code], https://doi.org/10.5281/zenodo.15628903

Vautard, R, Cattiaux, J, Yiou, P, et al. (2010). Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness. Nature Geoscience, 3, 756–761. https://doi.org/10.1038/ngeo979

Wang, K and G D Clow (2020) The Diurnal Temperature Range in CMIP6 Models: Climatology, Variability, and Evolution. Journal of Climate, 33, 8261–8279. https://doi.org/10.1175/JCLI-D-19-0897.1

Wedi, N, Sandu, I, Bauer, P, et al. (2025) Implementing digital twin technology of the earth system in Destination Earth. Journal of the European Meteorological Society, 3, 100015. https://doi.org/10.1016/j.jemets.2025.100015

WMO (2025) https://wmo.int/activities/group-earth-observations-geo/

Zha, J, Shen, C, Wu, J, Zhao, D, Fan, W, Jiang, H and T Zhao (2023) Evaluation and Projection of Changes in Daily Maximum Wind Speed over China Based on CMIP6. Journal of Climate, 36, 1503–1520. https://doi.org/10.1175/JCLI-D-22-0193.1

Zhang, C, Golaz, J-C, Forsyth, R, et al. (2022) The E3SM Diagnostics Package (E3SM Diags v2.7): a Python-based diagnostics package for Earth system model evaluation. Geoscientific Model Development, 15, 9031–9056, https://doi.org/10.5194/gmd-15-9031-2022

Zhang, Z and K Wang (2024) Quantify uncertainty in historical simulation and future projection of surface wind speed over global land and ocean. Environmental Research Letters, 19, 054029. https://doi.org/10.1088/1748-9326/ad3e8f