



Indicator Delta Scaling (IDS): Rethinking bias adjustment for climate risk indicators.

Jesús Peña-Izquierdo¹, Sascha Hofmann¹, Victor Estella-Perez¹, Tatiana Ray¹, Laia Romero¹ and Chiara Cagnazzo²

¹Lobelia Earth, Barcelona, 08003, Spain

²European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, 53175, Germany

Correspondence to: suso.pena@lobelia.earth

Abstract.

Translating global climate projections into local risk assessments relies heavily on Bias Adjustment (BA). Yet, standard BA methods often distort historical observational baselines and modeled trends, suffer from significant inconsistency across different regions, variables and processed datasets, while imposing heavy computational costs. Driven by the need for credible, consistent and computationally efficient climate information, we propose Indicator Delta Scaling (IDS), a unified framework that generalizes the traditional concept of Change Factors for direct application to any type of climate indicator. IDS relies on three core principles: securing accurate historical baselines by anchoring climate indicator estimates strictly to observations; generating bias-adjusted future indicators by directly scaling these baselines with raw climate model-based deltas; and shifting any required adjustment from the underlying input data to the indicator computation itself. Beyond its conceptual and physical simplicity, a systematic evaluation across a multi-model ensemble and a wide spectrum of climate indicators shows that IDS consistently outperforms state-of-the-art methods (QM, QDM, and ISIMIP3b) by reducing their error in adjusted future projections by 23% on average across indicators and climate models. At the same time, IDS provides notably superior consistency, shrinking error variability to one-third of that observed in standard methods. This performance gap stems primarily from the failure of traditional approaches to eliminate historical biases, leaving behind roughly one-third of the initial raw model error alongside severe local outliers, whereas IDS always ensures zero historical errors. Furthermore, while conventional methods introduce unpredictable artifacts, including artificial reversals in model trends and compounding errors, the skill and uncertainty of IDS remain strictly tied to climate model performance. Ultimately, by avoiding the BA of the daily variable distribution, IDS delivers notable computational speedups ranging from x1.2 to x12, depending on benchmark. Overall, IDS offers a transparent and computationally efficient framework that preserves observational baselines and model projected changes, while improving indicator level performance and consistency across the evaluated applications, making it a promising approach for large-scale climate indicator production and climate risk applications.



1 Introduction

Stakeholders across multiple economic sectors increasingly require ready-to-use and reliable climate information to support climate change adaptation and risk-informed decision-making. For these applications, it is essential that climate estimates are as realistic and precise as possible, characterizing accurately average conditions and complex climate extremes for both the historical baseline and future periods that underpin climate risk assessments. Bias adjustment (BA) methods represent a key processing step in the production of climate information derived from climate projections, aiming to reduce systematic model errors in selected climate models statistics and enhance the usability of climate simulations (Maraun, 2016).

Many studies have, however, demonstrated that most commonly used BA approaches introduce severe physical and mathematical inconsistencies. These include the undesired alteration of known observed historical estimates, the distortion or even reversal of the climate change signal projected by climate models, artificial changes in the model uncertainty spread, and a strong sensitivity of BA method's performance to the considered variable, climate indicator, region, and observational reference dataset (Casanueva et al., 2020; Chandel et al., 2024; Ehret et al., 2012; Iturbide et al., 2022; Lafon et al., 2013; Maraun, 2016; Padulano et al., 2025; Spuler et al., 2024). These limitations highlight the profound risks of applying BA techniques blindly as generic, black-box operations, without careful examination of their implications for each specific case. To address this, several publications have recommended the use of case-specific BA methods that aim to at least mitigate these issues for the specific statistical properties or physical processes relevant to a given application, though at the cost of methodological generality (Addor et al., 2016; Maraun, 2016; Maraun et al., 2017; Olschewski et al., 2023; Tootoonchi et al., 2023). However, these two contrasting paradigms create a critical bottleneck: the consistency approach is broadly applicable but often unreliable, while the use-case-specific approach is more physically robust but lacks generalizability (Spuler et al., 2026). Consequently, neither strategy fully satisfies the strong and urgent demand for a comprehensive provision of consistent, diverse and accurate global climate information required to support informed decision-making across sectors and stakeholders.

A radical alternative approach involves the full avoidance of BA to focus on the analysis of raw model-projected trends and anomalies alone (i.e. the difference between future and historical estimates) to identify areas more exposed to substantial changes in the future (Spuler et al., 2026). A clear limitation of this simplistic approach is, however, the lack of realistic absolute values that are indispensable for any climate impact and risk assessment. A hybrid approach that aims to exploit the benefits of working directly with intact modeled trends under a no-BA post-processing while still producing calibrated future estimates is the use of Change Factors (Anandhi et al., 2011). In its simplest form, more commonly known as the Delta approach (Hay et al., 2000; Maraun, 2016; Räisänen and Rätty, 2013), change factors simply represent the projected changes between future and historical climatological means and they are used to adjust the observational historical record producing future time series with bias-corrected climatological means. It is clear, however, that while this approach fully preserves projected changes in mean fields, any other modeled change signals beyond the mean are lost, as the resulting future



temperatures are simply an offset replica of the historical observations. A more effective version of this approach has typically been used in hydrology and urban drainage studies, where the change factors are computed from other aggregated derived indicators, not to adjust the continuous observational time series, but rather the corresponding observational indicator itself (Arnbjerg-Nielsen, 2012; Olsson et al., 2009; Willems and Vrac, 2011). (Willems, 2013), for instance, proposed extracting change factors for high percentiles or return levels of precipitation accumulations directly from raw climate projections. These are then directly applied to the corresponding observation-based estimates to obtain a model- and observation-consistent calibrated future indicator. However, the traditional Change Factor approach fails when computing threshold-based indicators (e.g., estimating the number of days exceeding a specific value), because these metrics require a bias-free time series to be valid. This specific requirement highlights, in fact, a broader distinction in the type of climate information ultimately required in different risk assessments, specifically, between the need of continuous bias-adjusted time series or calibrated derived climate indicators. While complex dynamic impact models require time series, climate indicators constitute by far the most commonly consumed format. By capturing the intensity, duration, or frequency of specific climate processes, aggregated climate indicators drive a much broader range of sectoral risk assessments.

Here, we propose, to the best of our knowledge, the first generalization and comprehensive evaluation of the Change Factor BA paradigm directly applied to any type of climate indicator. The presented Indicator Delta Scaling (IDS) method operates at the indicator level by first anchoring historical indicator estimates to the observational baseline and then scaling this baseline directly using the indicator's raw model-based projected trend. IDS guarantees valid results for any type of indicator by adding indicator-specific adjustments to the indicator implementation itself, rather than distorting the underlying input data. To benchmark the framework, an exhaustive analysis is carried out based on eight highly diverse types of indicators computed on a CMIP6 multi-model ensemble of historical simulations, using the ERA5 reanalysis (Hersbach et al., 2020) as an observational reference against which the performance of IDS is analyzed and compared with three state-of-the-art BA methods. Comparing the different methods across the historical and future periods, together with their ability to reproduce modeled and observed trends, allows for a complete understanding of the benefits and limitations of the distinct approaches. The results demonstrate that prioritizing the integral preservation of the two independent sources of climate information, observations and climate projections, does not require sacrificing predictive skill; rather, IDS significantly outperforms standard methods across all indicators. In response to concerns highlighted in recent literature (Maraun, 2016; Spuler et al., 2026), IDS represents a paradigm shift in the role of the BA procedures: rather than attempting to physically compensate deficiencies in climate models, it provides a transparent, uncorrupted bridge between historical observations and climate projections that enhances the usability of climate information. The development of the IDS framework was supported by the European Centre for Medium-Range Weather Forecasts (ECMWF) to generate a new extensive and diverse dataset of cross-sectoral climate indicators for the Copernicus Climate Change Service (C3S), developed to address the climate risk assessment requirements of the European Investment Bank (EIB) for the EU hazard taxonomy regulation.



The paper is structured into four additional sections. Section 2 introduces the IDS methodology in full detail, including possible edge cases not considered in the present study. Section 3 describes the selected datasets, BA methods, and climate indicators together with an explanation of the evaluation experiment and corresponding metrics. Section 4 presents all the results, describing the performance of the different methods across diverse and complementary dimensions, ranging from local, climate model, and indicator-specific analyses to fully aggregated method evaluations covering the four components: historical, future, modeled, and observed trends. The paper closes with the Conclusion section, where a general review of the presented results is discussed, establishing clear relations to the well-known problems presented recently in the literature and demonstrating how the IDS addresses many of these issues.

2 Indicator Delta Scaling (IDS)

Two independent and essential sources of information must be combined to generate calibrated future climate information. On the one hand, high-quality observations offer the most accurate representation of the historical climate and should always be preserved as a solid baseline against which future changes are compared. On the other hand, future changes in the climate system are provided by climate projections, which should likewise be preserved as the sole source of information regarding the climate change signal, representing the primary value of climate modelling.

The main goal of the Indicator Delta Scaling is to uphold these principles at every level when computing climate indicators. Furthermore, given the growing concerns highlighted in recent literature regarding common Bias Adjustment practices (Maraun et al., 2017; Spuler et al., 2026), a viable solution must also rely on a simple, transparent procedure. Its logic and impact on the final outputs should remain clear, controlled, and easy to interpret, avoiding the introduction of unexpected artifacts. Finally, to encourage broad adoption across the scientific community and climate services, the method should be computationally efficient, lightweight, and free of unnecessary overhead.

A straightforward conceptual solution to these requirements is the already mentioned Change Factor approach, where calibrated future estimates are calculated by applying the model's climate change signal directly to observational data. Beyond the standard formulations that work with the delta change of climatological means only (Hay et al., 2000) or the change factors for a very concrete derived metric (Olsson et al., 2009), IDS represents a generalization of this concept to any type of climate indicator. The procedure, sketched in Fig.1, follows three simple steps: 1) computing the corresponding climate indicator from both the historical observational reference and raw model data across historical and future periods, 2) estimating the delta change factors projected by the raw climate model and 3) scaling the observation-based historical indicator with these change factors to obtain the bias-corrected indicator for the future period. Because different indicator types involve different mathematical structures and physical processes, our approach requires checking whether an indicator needs specific adjustments to ensure consistent and reliable results, even when working with biased inputs. In essence, because each indicator

extracts only specific statistical or temporal characteristics of the underlying climate variable, instead of a generic adjustment applied broadly to the raw model data distribution, we propose applying required adjustments to each specific indicator, if needed, ensuring full consistency with the observational reference and project model change signal.

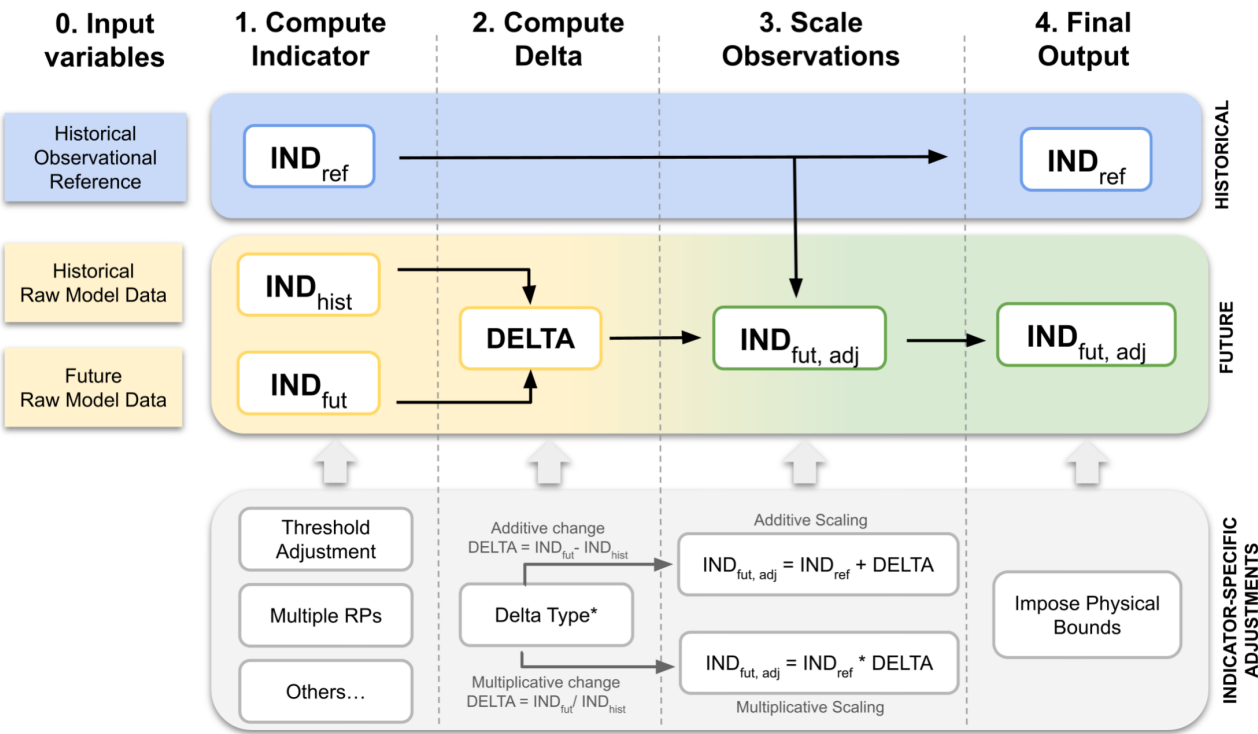


Figure 1. Schematic description of the Indicator Delta Scaling procedure where observational data is directly used to compute historical indicators (top blue panel). Raw climate model data is processed to compute corresponding historical and future indicators as well as the corresponding delta change factor (yellow part of the middle panel). Purely observational and modeled estimates are combined to produce bias-adjusted future indicators (green part of the middle panel). Indicator-specific adjustments for each of the steps are described in the bottom grey panel. *Types of delta change factor might also include other more complex formulations such as hybrid or bounded deltas, see text.

The first and systematic adjustment is deciding if the delta change factor of a specific indicator is better represented by an additive or a multiplicative change factor. This is a non-trivial question with likely no universal answer. As a general rule, we use multiplicative change factors for zero-based indicators with an open upper bound (such as RX1day; the maximum daily precipitation accumulation). Additive deltas are systematically used with interval variables or two-sided bounded indicators



(such as TXx; maximum daily maximum temperature or TX30; number of days above 30°C). See Table 1 to check the delta factor used for each indicator. Although more sophisticated hybrid or bounded delta formulations exist (see (Lange, 2019) to prevent division by near-zero denominators or out-of-range values in bounded indicators, we intentionally restrict this study to strictly additive and multiplicative factors to demonstrate that maximizing the preservation of the original model change signal performs effectively across global domains. Nevertheless, advanced delta types remain a practical optional refinement that IDS users can select for specific edge cases.

The most obvious adjustment when working with biased model data applies to threshold-based climate indicators (e.g., TX30 or CDD, the number of consecutive dry days). Again, rather than bias-adjusting daily climate data and applying a fixed threshold everywhere, IDS shifts perspective by adjusting the threshold itself for each model and location. Following (Hoffmann et al., 2018), we use a quantile-based approach: first, we determine the observational quantile corresponding to the target threshold at each location (30°C for TX30 or 1mm/day for instance for CDD); second, we extract the climate model local thresholds corresponding to those same quantiles from the historical simulation; and third, we use these spatially adjusted thresholds to compute the indicator across both historical and future periods. For threshold values falling outside the observational range, we offset the observational min/max by the difference from the original threshold. This method essentially simplifies standard Quantile Mapping by evaluating only the specific quantile associated with the fixed threshold.

Other types of adjustments might be required on very specific indicators. As indicated above, the strict combination of independent observational baselines and model deltas can cause out-of-range values in bounded indicators. For instance, for indicator TX30, the number of days in a year must always lie within the interval [0, 360] for a ‘360-day’ calendar. Under specific conditions, IDS additive deltas can produce physically impossible values. To resolve this while preserving the exact model delta across valid ranges, the implementation of the IDS in this study applies physical boundary enforcement (hard clipping) as a final post-processing step. In specific cases, however, users might prefer using the bounded delta formulations mentioned earlier (Lange, 2019) that smoothly ensure correct bounded outputs while slightly modifying projected change signals. Another edge case, not included in this paper, occurs when computing extreme value indicators across multiple return periods. Computing return levels first to later obtain corresponding model deltas can lead to non-monotonic final return levels when scaling observations. This typically occurs when models project larger deltas for shorter return periods than for longer ones, combined with a relatively flat observational return level plot. Following a procedure similar to (Willems and Vrac, 2011), this can be resolved by: 1) extracting annual maxima for observations alongside historical and future model runs, 2) ranking all annual maxima series, 3) calculating rank-specific deltas on the model data, 4) applying these deltas to the ranked observational maximas, and 5) applying Extreme Value Analysis to the scaled series to guarantee strictly monotonic return level outputs.



Finally, for indicators relying on complex highly non-linear formulations (e.g., FWI90p), input variable biases would expectedly propagate into strongly biased final outputs and correspondingly on biased delta signals. To mitigate this issue in IDS, we tested a linear mean-scaling preprocessing step on input variables, adjusting model climatological means to match observational reference before computing the indicators. However, this step yielded no consistent or noticeable performance gains overall for indicators considered in this study, so it is omitted by default. Yet, users may consider pre-scaling input variables for certain complex indicators to reduce mean input biases while still preserving projected delta change signals

3 Data & Methods

3.1 Data

To evaluate the different BA methods, following Spuler et al. 2024, we use historical climate projections together with overlapping observational records to test the performance of bias corrected simulations in representing diverse aspects of the climate change signal. In this study, the spatial domain is global and covers the period 1951-2010. For the evaluation, the timeframe is split into training (1951 to 1980) and validation period (1981 to 2010). However, and for clarity, we use the terms historical and future periods throughout the text to refer correspondingly to these respective periods. Although the validation period also falls within the historical record, in practical terms it effectively serves as an out-of-sample period analogous to a future projection period that can be evaluated.

The observational reference is represented by the ERA5 reanalysis (Hersbach et al., 2020), while historical climate projections come from six CMIP6 models with realistic climate sensitivities (Hausfather et al., 2022): CMCC-ESM2, FGOALS-G3, GFDL-ESM4, MIROC6, MPI-ESM1-2-LR, MRI-ESM2-0. The spatial resolution of both datasets, ERA5 and CMIP6 models, has been homogenized to a common 4°x4° grid, limiting the global domain to latitudes 60°S–70°N and longitudes 150°W–160°E: First ERA5 0.25° grid resolution is coarsened to a 1° degree and now, with a similar resolution to CMIP6, both datasets are linearly resampled into the common grid. The goal of these geographical decisions is to reduce computational costs while having no impact on the main conclusions of this study.

3.2 Reference BA methods

We consider 3 widely used Bias Adjustment methods to benchmark the IDS. The non-parametric Quantile Mapping (QM, (Deque, 2007) represents a mid-complexity method that aims to adjust the modeled climate projection distribution to match the observed historical distribution. A major limitation of the QM method is that it does not impose any condition to preserve the change signal projected by the underlying climate model, which can lead to artificial and physically unjustified trends (Cannon et al., 2015). A major improvement on this type of approach was provided by Cannon et al., 2015) with the Quantile Delta Mapping (QDM) which is a simple yet effective modification of the standard QM. While historical bias correction



replicates standard QM output, QDM applies quantile-specific delta change factors to future projections to ensure modeled climate trends are explicitly preserved at each quantile. We follow Cannon et al., 2015) procedure for the correction of dry/wet frequency in precipitation for both the QDM and QM implementation: Values smaller than 0.05 mm/day are substituted by randomized values from a uniform distribution between 0 and 0.05 mm/day prior bias correction in both observational reference and model data. Later, bias corrected output values below 0.05 mm/day are substituted by 0 mm/day. Finally, we consider the ISIMIP3b method (Lange, 2019), which serves as the official data pipeline for the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). ISIMIP3b represents a complex, variable-specific bias correction framework based primarily on parametric quantile mapping. It enhances multi-variable consistency, enforces physical range boundaries, and smoothes extreme tails using theoretical distributions.

While these three methods differ in their complexity, they all share one critical feature: modifying modeled time series to match historical observational statistics. Although this adjustment significantly reduces model biases, it can also alter temporal and distributional characteristics of the model data that are relevant for derived climate indicators, such as heatwave durations, drought intensities or return periods extreme values.

3.3 Climate indicators

A set of eight distinct indicator types (Table 1) was selected to evaluate IDS against state-of-the-art BA methods. Each indicator represents an individual climate phenomenon or the combination of phenomena (including process intensity, threshold-based conditions, event frequency, spell duration, variability ranges and high non-linear formulations) covering both acute and chronic hazards using single or multiple essential climate variables. For indicator TX30, the evaluation domain has been limited to land to obtain more meaningful results since almost no ocean sites exceed 30°C temperatures. As described in Section 2, IDS represents an indicator-specific bias adjustment approach, implying that certain indicators may require tailored adaptations. Nevertheless, we consider the selected set of indicators to be sufficiently representative of most commonly used climate indicators. The selected indicators have been chosen not only to span different hazard types and formulations, but also to sample distinct statistical characteristics of the underlying climate variables, including mean state, threshold exceedances, persistence, distribution tails, and extreme-value return levels.

Table 1. Description of the different climate indicators used in the evaluation including the type of hazards and mathematical characteristics of the corresponding formulation.

Short Name	Description	Type	Input Variables	Units	Delta Type	Bounds
TXx	Average maximum daily maximum temperature	Process intensity, Acute	Tasmax	°C	Additive	-



TDRx	Average maximum temperature diurnal range obtained as the difference between daily maximum and minimum temperature	Process intensity, Multi-variable, Acute	Tasmax, Tasmin	°C	Additive	-
TX30	Average number of days in a year with daily maximum temperature above 30°C	Threshold-based, Chronic	Tasmax	day	Additive	[0, 360]
CoolDD	Average Cooling Degree Days, following Spinoni et al. (2018), quantifies the excess of heat in degrees above specific thresholds across daily minimum, mean and maximum temperature	Threshold-based, Variability range, Multi-variable, Chronic	Tasmin, Tas, Tasmax	Degree days	Additive	[0, -]
CDD	Average number of Consecutive Dry Days, defining dry day one with daily precipitation accumulation below 1mm/day	Threshold-based, Spell duration, Chronic	Pr	Day	Additive	[0, 360]
PRCPTOT	Average total precipitation accumulation considering wet days (daily precipitation threshold 1mm/day)	Process intensity, Threshold-based, Chronic,	Pr	mm/day	Multiplicative	-
RX1day_20yr	Precipitation accumulation in a day for a 20 year return period event. A GEV distribution is fitted to climate period annual maxima time series through the PWM method.	Process intensity, Event frequency, Acute	Pr	mm/day	Multiplicative	-
FWI90p	Average percentile 90 th of the Fire Weather Index (FWI). The FWI is a wildfire risk indicator purely based on weather conditions	Process intensity, High non-linear, Multi-variable, Acute	Tas, Pr, sfcWind, Hurs	-	Additive	[0, -]

3.4 Evaluation approach

A key novelty of our evaluation approach compared to (Spuler et al., 2024) and previous studies (Casanueva et al., 2020) is the assessment of all relevant components affecting the accuracy of bias-corrected future climate indicators. This includes not only assessing the preservation of the projected change signal and the accuracy of future period estimates but, equally important, the preservation of observational estimates during the historical period alongside the actual observed change signal. This addition is critical because BA methods might introduce errors into the historical representation of indicators compared to the observational reference. These errors can later be offset by additional errors in projected trends, ultimately leading to a future performance that looks better than it actually is. Moreover, errors in the representation of future period indicators do not depend only on the performance of the BA method, but also heavily on the skill of the climate models themselves, alongside the influence of natural variability. Although historical simulations used in this study are uninitialized, and thus are not synchronized with observed internal variability, the proposed evaluation setup is fully valid because our objective is not to quantify the absolute accuracy of the prediction system, but rather to isolate and compare how different BA methods transform both historical baselines and projected changes. Evaluating these involved components independently helps to disentangle the



actual added value of each BA method. Taking all of this into account, we evaluate errors across four main components: 1) historical estimates compared to the observational reference, 2) representation of observed trends, 3) the preservation of projected trends relative to raw climate projections, and 4) future estimates compared to the observational reference.

To do so, we consider four different metrics. Errors for the historical and future periods are always presented in absolute terms as the difference between the estimate and the observational ground truth at each location for every model. Changes, however, are expressed in either absolute or relative terms (i.e., percentage change relative to the historical baseline) depending on the indicator (Table 1). Correspondingly, their errors are also represented in absolute or relative (%) terms. To compare performance across diverse indicators, we normalize the errors by dividing them by a representative reference value. For absolute errors, this reference is the corresponding climatological Global Spatial Variability (cGSV), defined as the standard deviation across all grid locations during the historical period. Thus, a normalized error greater than 1, or 100% expressed as a percentage, means the local error exceeds the indicator's global spatial variability. For relative errors, the reference value is the historical cGSV divided by the indicator's historical global mean. We then aggregate these normalized location-based errors per model by computing the Normalized Root Mean Squared Error (*NRMSE*) for a global summary. Finally, we evaluate the *NRMSE Skill Score* (*NRMSESS*), which directly compares the error of a given BA method against the biased raw model error:

$$NRMSE_{BA} = 1 - \frac{NRMSE_{BA}}{NRMSE_{RAW}} \quad (1)$$

A *NRMSESS* above 0 means the BA method alone improves raw model skill, while a negative value shows it degrades skill. A perfect score of 1 occurs only when $NRMSE_{BA}$ reaches zero, i.e. the bias-adjusted estimate perfectly reproduces the ground truth. In short, the *NRMSESS* represents the proportion of the initial error in the raw climate model that the corresponding BA method removes.

4 Results

4.1 Evaluation at local level

To evaluate IDS against standard BA methods, we begin with a representative example: the extreme temperature indicator, TXx, for the CMCC-ESM2 model. As established in Section 3.4, relying on uninitialized historical simulations, despite their inherent deviation from observed internal variability, which contributes to larger prediction errors, still allows us to fully isolate how each BA method transforms the different components of the climate model data. Figure 1 shows how the raw model run and its bias-adjusted counterparts capture the historical (1951–1980) and future, i.e. pseudo-future, period (1981–2010) compared to the ERA5 reference.



283
 284 The CMCC-ESM2 historical simulations display severe cold and warm biases (Fig. 2b,h). The largest warm biases relative to
 285 historical observations, reaching above +3°C, are concentrated in continental regions such as Europe, Western Asia, and central
 286 North and South America. Conversely, the strongest cold biases are located primarily over oceans, especially in the North
 287 Atlantic, Eastern Indian, and South Pacific basins. On the other hand, observed trends over the 60-year period show widespread
 288 global warming, with a few regional hotspots like Europe and Northern Canada already experiencing a warming of +2°C (Fig.
 289 2m). Some regions also present localized, minor cooling, such as central South and North America. CMCC-ESM2 replicates
 290 the observed global warming pattern, albeit with certain regional differences (Fig. 2s, n), for instance, a weaker warming over
 291 most of Europe and northern North America but an overestimation of future extreme temperatures in central South and North
 292 America.

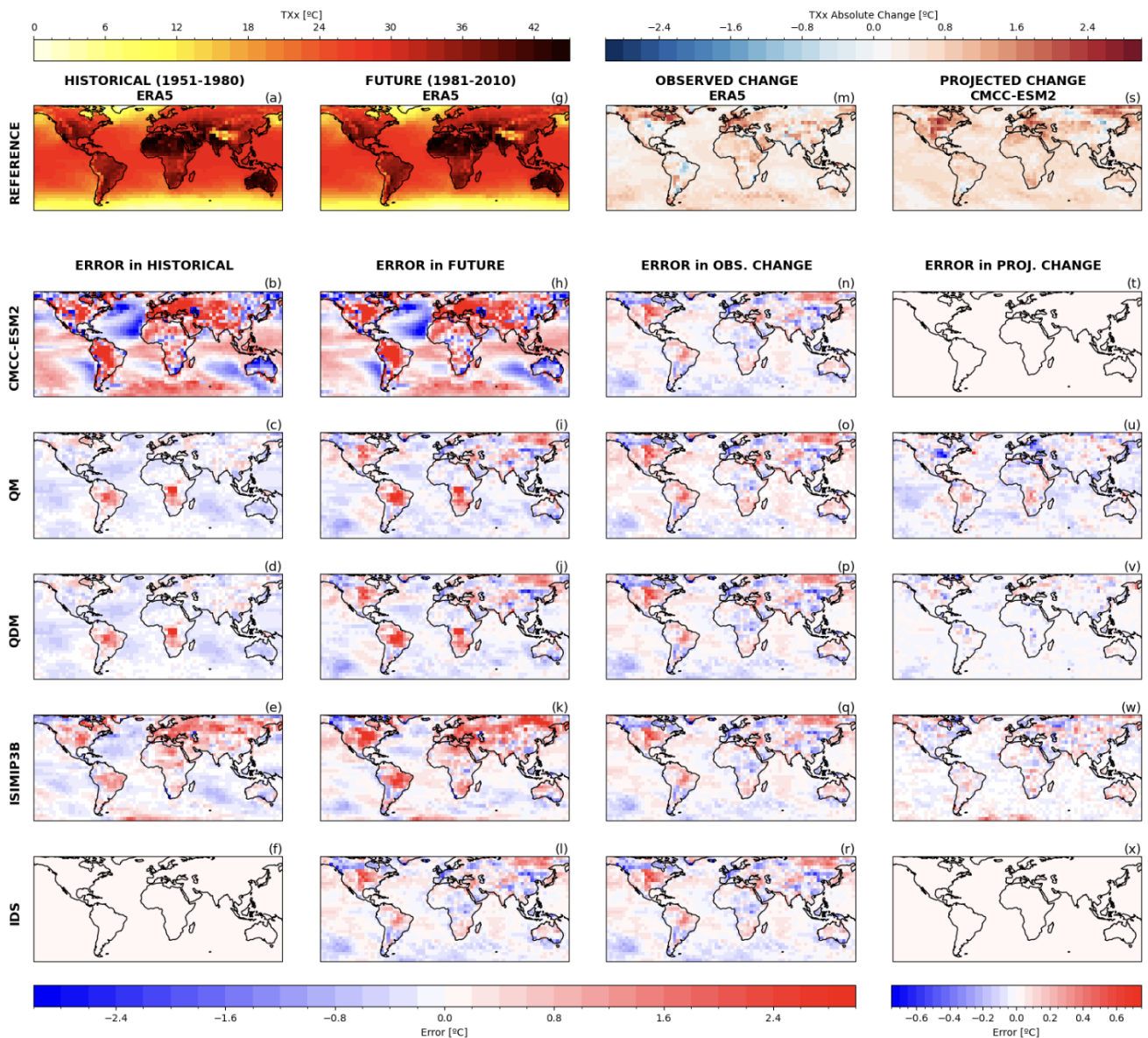
293 Regarding the performance of the BA methods, we first examine the historical period (Fig. 2c-f). Standard methods already
 294 display striking errors relative to the observational reference, with significant differences also emerging between the individual
 295 methods. QM and QDM show the same pattern, since they differ only in their procedure for estimating future periods, with
 296 warm biases exceeding +1°C remaining in central South America and South Africa, the latter not even present in the raw
 297 historical simulation. ISIMIP3B displays more homogeneous biases, also reaching around +1°C in several regions; it
 298 reproduces spatial patterns similar to CMCC-ESM2, but with smaller errors. IDS, in stark contrast, shows no error by
 299 construction because it simply replicates observations.

300 Trend distortions introduced by the BA methods are considerably smaller (up to four times lower) than baseline historical
 301 biases (Fig. 2u-x), showing that all methods acceptably preserve the raw CMCC-ESM2 change signal with deviations typically
 302 well below 0.5°C. However, distinct differences between methods re-emerge, now also between QM and QDM, where the
 303 latter exhibits, as expected, the smallest errors, close to 0 in most regions. QM modifies projected trends, particularly in Eastern
 304 Europe and central US, leading to cooler-than-modeled changes. Differences in ISIMIP3B appear less well defined and more
 305 scattered overall but still not negligible. Once again, IDS replicates by construction the delta change from the raw model and
 306 thus exhibits no error.

307 The strong preservation of modeled trends across these four methods leads to an almost identical error representation of
 308 observed trends also when compared to raw CMCC-ESM2 (Fig. 2n-r). However, it is the differing performance during the
 309 historical period that explains why errors in the future period diverge across methods (Fig. 2h-l). First, differences are minimal
 310 between QM and QDM, arising solely from the small modification of projected trends by both methods as previously noted.
 311 The substantial remaining biases in the historical representation of ISIMIP3B, combined with its non-negligible alteration of
 312 the projected change, lead to large errors in future estimates across regions similar to those of the raw climate simulation. A
 313 revealing case of error interaction occurs in central Europe, where ISIMIP3B presents large warm historical biases around
 314 +1°C alongside a preserved colder-than-observed trend projected by the CMCC-ESM2 model. These two opposite errors end



up cancelling out when combined for the future period estimate, yielding almost no error. However, this accurate estimate is merely a mathematical coincidence rather than a true physical representation of projected processes or physically corrected biases. Once more in contrast, the error in the future estimate for IDS is straightforward: it simply equals the error of the raw model's projected change owing to its exact representation of both historical observations and modeled deltas, thereby ensuring clear interpretability and confidence in the outcomes.



320



Figure 2. Spatial evaluation of the TXx indicator [$^{\circ}\text{C}$] for historical (1951–1980) and future (1981–2010) periods. The top row shows ERA5 reference historical (a) and future (g) representations together with the corresponding observed changes (m) and raw CMCC-ESM2 model projected changes (s). Corresponding absolute differences [$^{\circ}\text{C}$] for raw CMCC-ESM2 model (b,h, n, t), QM (c, i, o, u), QDM (d, j, p, v), ISMIP3B (e, k, q, w) and IDS (f, l, r, x) bias-adjusted outputs relative to respective reference historical bias are shown.

Extending this same analysis to all other evaluated indicators reveals a similar pattern throughout, with important differences depending on location, indicator, method and climate model (Figs. 3 and 4. See also supplementary material Figs. S1 to S8 for map versions for each indicator). The first striking result from Figure 3 is that, in general, standard BA methods are unable to fully remove biases in historical estimates, ranging from 5–10% (PRCPTOT, TXx) to 30–50% (FWI90p, CDD, DTRx) of the cGSV, which is notably concerning. A particularly worrying case arises for the bias-adjusted, frequency-based indicator RX1day_20yr, which shows a systematic underestimation relative to the cGSV of 10% for QM and QDM and 25% for ISIMIP3B. In some regions, this underestimation exceeds 50 mm/day, representing approximately 150% of the cGSV, raising serious doubts about the practical use of these standard methods for specific indicators. The only exceptions occur with QM and QDM for the threshold-based indicators TX30 and CoolDD, where historical errors are virtually negligible. These two methods also consistently center errors around 0 across indicators, with the exception of RX1day_20yr. In contrast, ISIMIP3B clearly struggles more to systematically avoid both over- or underestimation of the historical reference. This is likely due to the use of parametric distributions in the quantile mapping correction of ISIMIP3B which makes a perfect fit to the data difficult. It is important to note that beyond the bulk of the distribution, specific locations present extreme outliers across most indicators with errors exceeding 50%, or even 100%, of the cGSV even for the training and historical period (Fig. S9). Future estimates display error ranges similar to that of historical estimates across most indicators, with the exception of PRCPTOT and CoolDD which show small errors for the historical. This similar error range is intriguing considering the additional error involved in future estimates linked to the skill on the predicted change signal. In this line, the IDS presents systematically smaller errors for the future, as no error is introduced during the historical period. A specific notable case for the IDS is the FWI90p indicator, which relies on a highly non-linear parameterization of four interacting variables. Strikingly, computing the delta change using raw, uncalibrated inputs and applying it to the observational baseline produces future estimates with lower errors than standard BA methods that explicitly calibrate those inputs (Fig. 3h). This demonstrates the profound advantage of anchoring to a perfect historical baseline, indicating that errors inherent to the uncalibrated delta change have a much weaker impact on final future estimates than the residual historical biases left by standard methods. We return to the broader significance of this finding in the discussion.

Regarding the distortion of projected climate change signals, a distinction must be made for threshold-based indicators. Because shifting the historical distribution directly alters threshold exceedance rates, any bias adjustment of these indicators (TX30, CoolDD, CDD, and PRCPTOT) inevitably, and correctly, modifies the raw model trends. Consequently, such



modifications should not be interpreted as method failure. Nonetheless, deviations from raw projected changes are generally small across all indicators, though subtle distortions remain present across most indicator–method combinations, with QM and QDM applied to PRCPTOT and RX1day_20yr standing out as notable exceptions (Fig. 4). Once again, ISIMIP3B's handling of RX1day_20yr stands out as an exception, systematically amplifying trends on average by 20% and yielding errors that reach up to 60% of the cGSV. Because methods largely preserve the model's underlying change signal for most indicators, their performance in capturing observed trends is nearly identical, including IDS, effectively mirroring the inherent skill of the raw climate simulations. Remarkably, while errors in representing observed trends are expectedly larger than the distortions introduced into the change signal itself, the residual errors remaining in historical estimates after bias correction are of a similar order of magnitude. As a result, both error sources, in historical estimates and observed trends, contribute comparably to the total errors observed in future period estimates. Figures S9 and S10 replicate Figures 3 and 4 while including extreme outliers to provide a complete description of these issues.

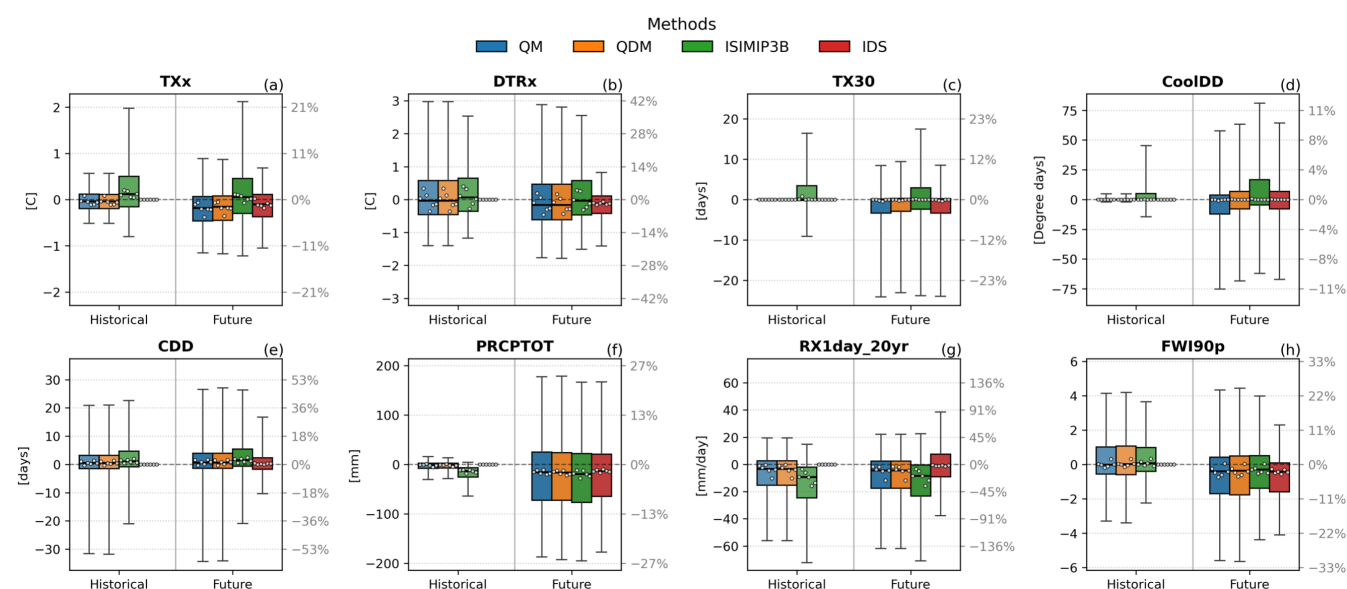


Figure 3. Absolute errors considering all locations and CMIP6 models between bias-adjusted estimates and observational reference for the four BA methods are shown for both historical (left) and future (right) periods across the eight climate indicators: TXx (a), DTRx (b), TX30 (c), CoolDD (d), CDD (e), PRCPTOT (f), RX1day_20yr (g) and FWI90p (h). Boxplots indicate the 25, 50 and 75th percentiles together with whiskers covering up to the 5 and 95th percentiles of the samples. Outliers are not shown, for this see Fig. S9. The right vertical axis shows the normalized error obtained by dividing the absolute error by the cGSV. Median error for each of the climate models is shown with a white dot.

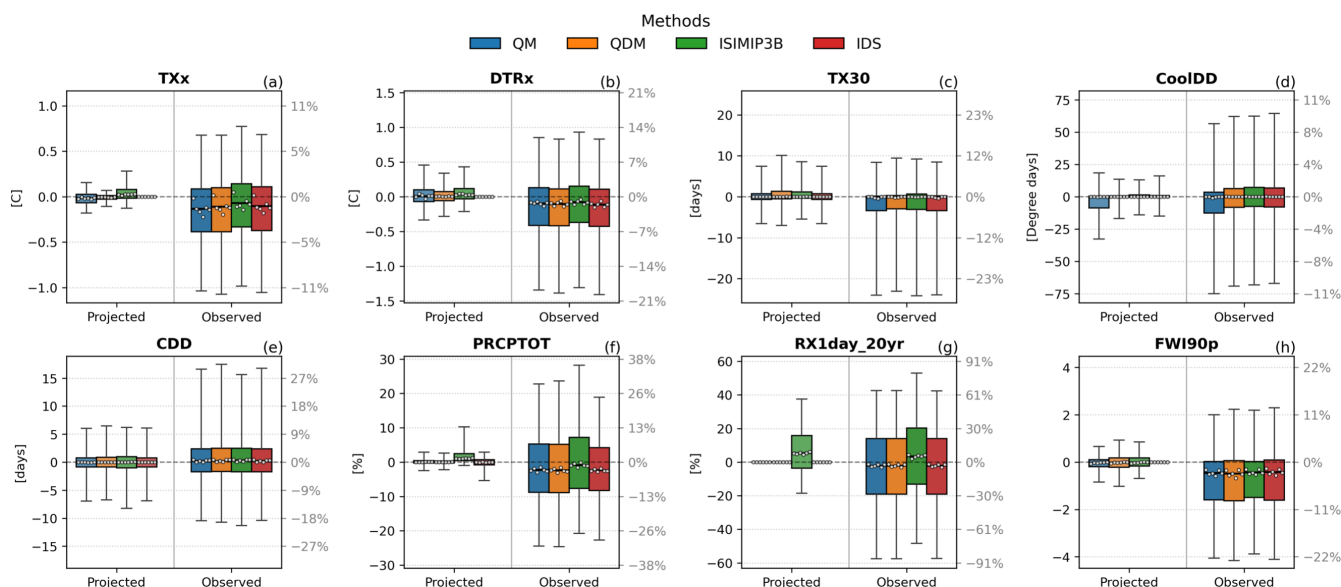


Figure 4. Absolute and relative errors [%] considering all locations and CMIP6 models between bias-adjusted estimates and model projected (left) / observed (right) trends are shown for the four BA methods across the eight climate indicators: TXx (a), DTRx (b), TX30 (c), CoolDD (d), CDD (e), PRCPTOT (f), RX1day_20yr (g) and FWI90p (h). Boxplots indicate the 25, 50 and 75th percentiles together with whiskers covering up to the 5 and 95th percentiles of the samples. Outliers are not shown, for this see Fig. S10. The right vertical axis shows the normalized error obtained by dividing the absolute error by the cGSV. See text for relative error normalization. Median error for each of the climate models is shown with a white dot.

4.2 Reversal of projected change

A critical issue regarding the modification of projected changes occurs when model trends are reversed after bias adjustment. Such reversals are particularly relevant for climate-service applications, as they can materially alter the projected climate-change signal delivered to users, while purely statistical BA methods generally provide no physical basis for interpreting the reversed signal as more plausible. To demonstrate the relevance of this issue, Figure 5 compares bias-corrected trends to raw climate model projections across three distinct indicators exhibiting both positive and negative trends.

Once again, different methodologies yield diverse results depending on the indicator. For DTRx and CDD, all standard methods produce strict trend reversals in 10% to 18% of all samples, with a "strict reversal" defined as a bias-adjusted change simply having the opposite sign of the modeled one. This proportion drops sharply to around 1% for significant reversals, where the shift originates from a non-negligible projected change (defined here as falling outside the central quintile). A special



case emerges for the RX1day_20yr indicator: while QM and QDM show negligible reversals, ISIMIP3B exhibits strict reversals in one-third of samples and significant reversals in 15% of samples with values significantly distinct from 0.

To evaluate whether these reversals are in some manner valid or justified, we compare the performance of the reversed samples in capturing historical observed trends against the full sample set (Fig. S11). The results indicate that samples with strict trend reversals show in general similar skill than the overall dataset for all three standard methods. However, for those cases with a significant reversal of the projected change, a noisier signal emerges with different patterns in the three methods. While QDM yields a skill reduction ($\text{NRMESS} < 0$) for most models in these samples, QM achieves the largest multi-model median skill gain, exceeding 0.5, though it still severely underperforms for certain models. Parametric ISIMIP3B, by contrast, presents a similar but highly robust skill gain across all indicators. However, because these significant reversals represent such a tiny fraction of samples ($\sim 1\%$, excluding RX1day_20yr at 14.6%), and because the more frequent strict reversals contrast these results which also aligns with the negligible skill observed across the complete data distribution beyond reversals, we cannot draw definitive conclusions from these localized skill improvements. In summary, the frequent occurrence of these reversals, without a systematic gain in skill, indicates they behave more like mathematical artifacts rather than true physical corrections, which poses serious risks for downstream climate applications.



Reversal of Projected Change

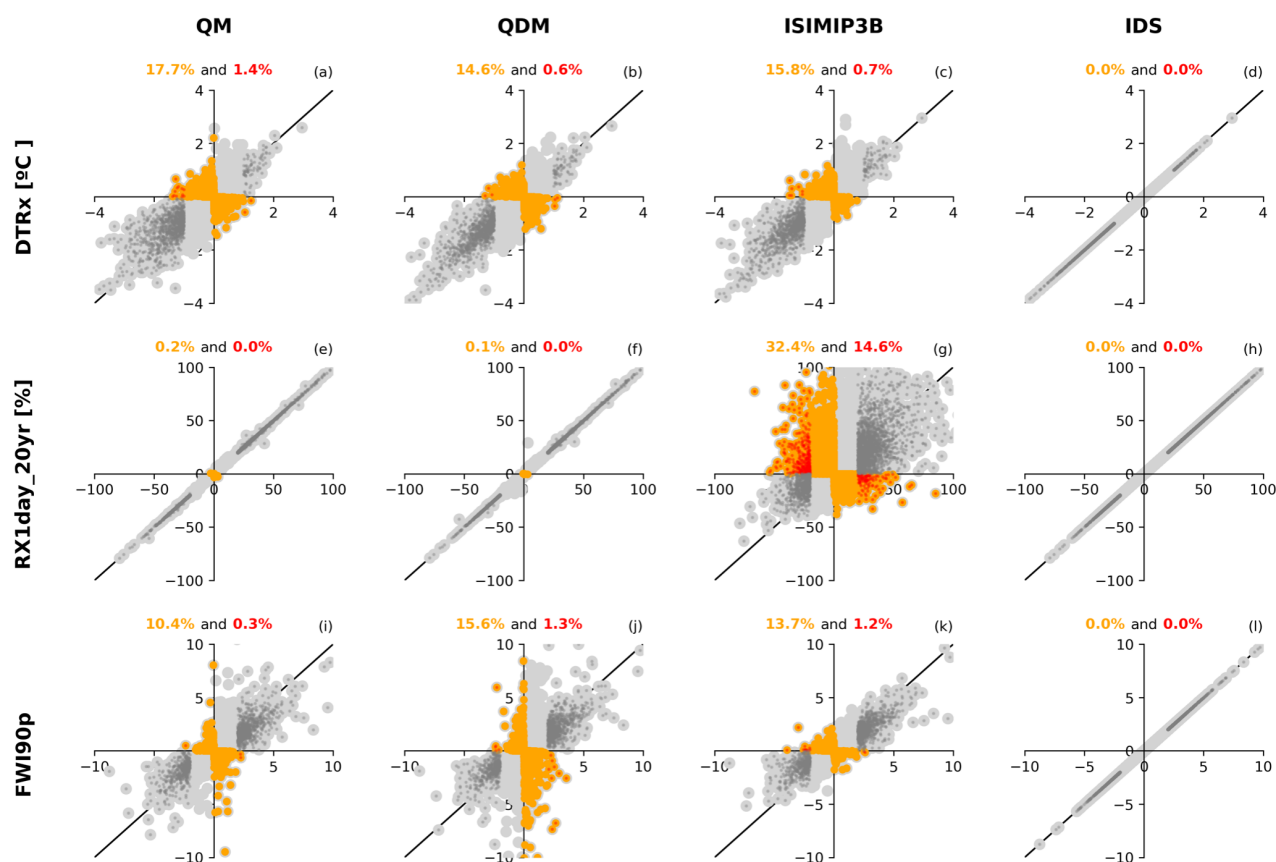


Figure 5. Scatter plots representing all locations for all CMIP6 models (light grey dots) observed changes (x-axis) against the bias-adjusted projected trends (y-axis) for DTRx (a-d), RX1day_20yr (e-h) and FWI90p (i-l) indicators for the four BA methods: QM (left), QDM (middle left), ISIMIP3B (middle right) and IDS (right). Samples with opposite signs (strict reversals) of projected trends relative to observations are shown with yellow dots. Samples with significant reversals are shown in red. Samples with correct trend signs out of the thresholds used for defining significant reversals are shown with dark grey dots. The percentage of yellow / red dots relative to the number of light grey / dark grey dots is shown for each case with yellow / red numbers. Threshold values used are $\pm 1^\circ\text{C}$ for DTRx, $\pm 20\%$ for RX1day_20yr and ± 2 for FWI90p

4.3 Evaluation at global level

To obtain a clearer general picture of the performance of the different methods, we aggregate errors by computing the NRMSE for each indicator and climate model (Fig. 6). Performance varies greatly, up to two orders of magnitude, across different dimensions. As also already revealed in Fig. 3, for the historical period significant errors remain, even though the three standard



BA methods significantly reduce climate model biases. After bias correction, the historical NRMSE averages 15% of the cGSV for empirical quantile-based methods, whereas it increases to 19% for ISIMIP3B. Two contrasting exceptions emerge: TX30, for which all QM and QDM completely eliminate model bias, and RX1day_20yr, which exhibits almost no error correction compared to the raw climate models in any of the methods and with NRMSE remaining on the order of 100%, i.e. errors after bias correction are as large as the observed global spatial variability of historical RX1day_20yr. IDS completely avoids this ineffective bias removal, consistently replicating observations for every indicator. Preservation of projected trends, only expected for non-threshold-based indicators, is mostly achieved with the clear exception again of RX1day_20yr in ISIMIP3B which yields an NRMSE of 30%. In general QM and QDM clearly outperform ISIMIP3B in the preservation of model signal, with the latter more than doubling the NRMSE of the former and reaching up to 10% of the cGSV. It is also interesting to note that although QM is not a trend-preserving method, it performs very similarly to QDM for these specific non-threshold-based indicators. Once again, IDS demonstrates full consistency with climate models, except, for threshold-based indicators where threshold adjustment is mandatory for valid results.

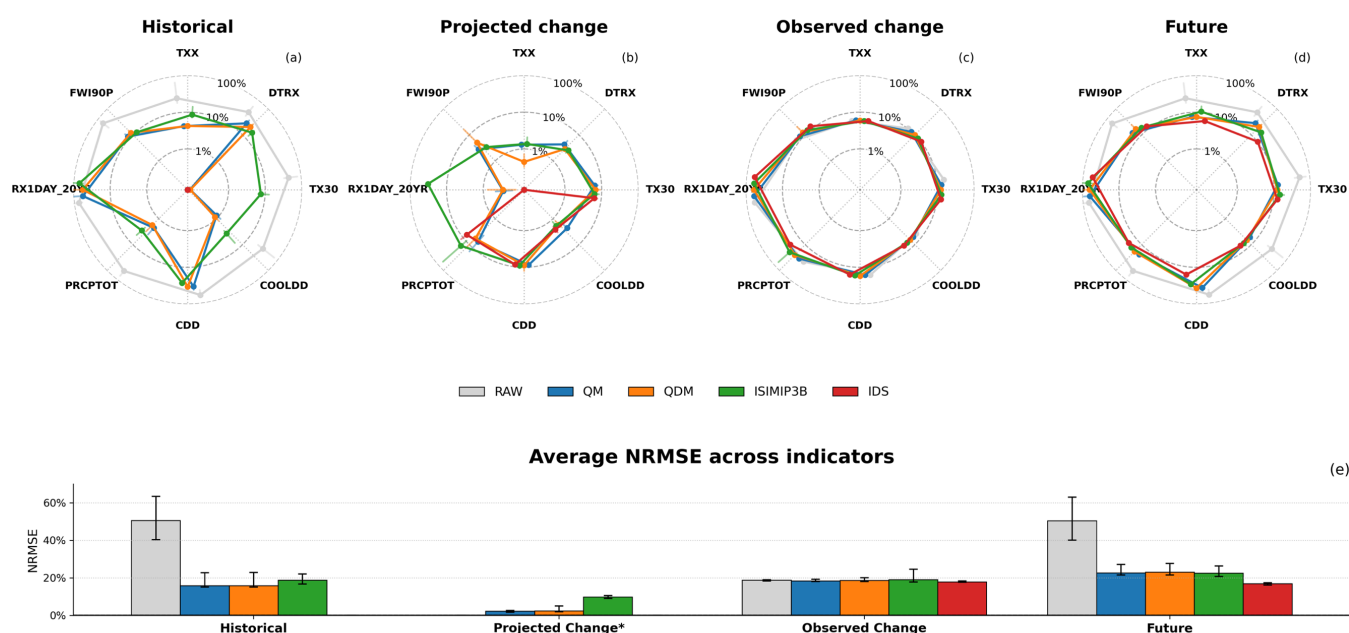


Figure 6. Top panel: Radial plots of the NRMSE for each of the eight indicators are shown for both the raw climate models (grey) and the four BA methods (in color) considering the Historical (a), Projected Change (b), Observed Change (c) and Future (d) estimates. The circular lines represent the median of the multimodel ensemble and the vertical radial segments the minimum and maximum error within the ensemble. Note the radial axis is in logarithmic scale. Bottom panel: Bar plots with the multimodel ensemble median NRMSE averaged across all indicators for the four considered estimates and four different



methods. Vertical segments represent the minimum and maximum error within the ensemble. *For Projected Change error, only non-threshold-based indicators are considered. Note that the smaller the NRMSE, the more accurate the method is.

Given the generally minimal trend distortion introduced by BA methods, errors in representing observed trends are virtually identical across methods and closely align with the performance of raw climate models. Performance in projecting trends therefore depends primarily on the intrinsic skill of the underlying climate model rather than on the chosen BA method. For future projections, however, errors stem from a combination of two a priori independent sources: climate model accuracy in representing observed trends and residual biases remaining in historical after bias correction. Nonetheless, QM, QDM, and ISIMIP3B exhibit a pattern of future errors across indicators that closely mirrors that of observed trends, achieving comparable mean future NRMSE around 22%, a substantial reduction compared to raw climate models (~50%) due to the mitigation of historical biases. IDS outperforms all candidate methods with a lower average future error (17% NRMSE, this is a 23% relative improvement compared to other methods), showing particularly sharp error reductions for indicators such as TXx, DTRx, and CDD.

An interesting feature becomes evident when examining the error propagation in Fig 6e: errors do not add linearly. According to the error propagation theory, errors of independent sources combine in quadrature (root sum of squares). For instance, in the raw model data, future errors are nearly identical to historical errors, even when trend errors are non-negligible, because the larger historical bias quadratically dominates the smaller trend error. For the standard BA methods, however, the combination of remaining historical biases and trend errors leads to even smaller total future error than quadratic propagation predicts. This yields a small gap relative to IDS that is somewhat artificial, given that IDS future error is governed by a single error source: the model's trend accuracy. To further examine this feature, we evaluate the possible presence of systematic error compensation between historical and trend errors. Note that we already described a clear example of error compensation in TXx (Fig. 2) for the ISIMIP3B method in Central Europe. Two sources of error combine independently (in quadrature) only when they are uncorrelated; if correlated, their combined error becomes smaller or larger than the expected quadrature sum depending on whether the correlation is negative or positive. Figure 7 shows that significant spatial anti-correlations exist across a large number of indicators, models, and BA methods, leading to substantial and systematic error compensation and thus a notable reduction in total future errors. The most pronounced case is once again the RX1day_20yr indicator, where the actual future error is 20% to 35% smaller than the independent combination of errors, presenting an anti-correlation ranging between -0.2 and -0.4. Other indicators exhibiting significant anti-correlation and error compensation are DTRx, CDD, and, for certain climate models, TX30, with error reductions on the order of 10%. There are also some cases featuring positive correlation, and thus constructive error propagation, producing larger than expected biases, although these appear only for isolated model-indicator-method combinations: for instance for CoolDD for one single model under QM method. Interestingly, all these correlations are already present in the raw model data and do not appear to be systematically amplified by BA methods. Nevertheless, this mechanism appears differently on different methods, models and indicators introducing unjustified



uncertainty into the final estimates. Regardless of the underlying cause, the presence of this fortuitous error interference represents an additional vulnerability that the direct approach of the IDS avoids entirely having one single source of error, model skill in predicting trends.

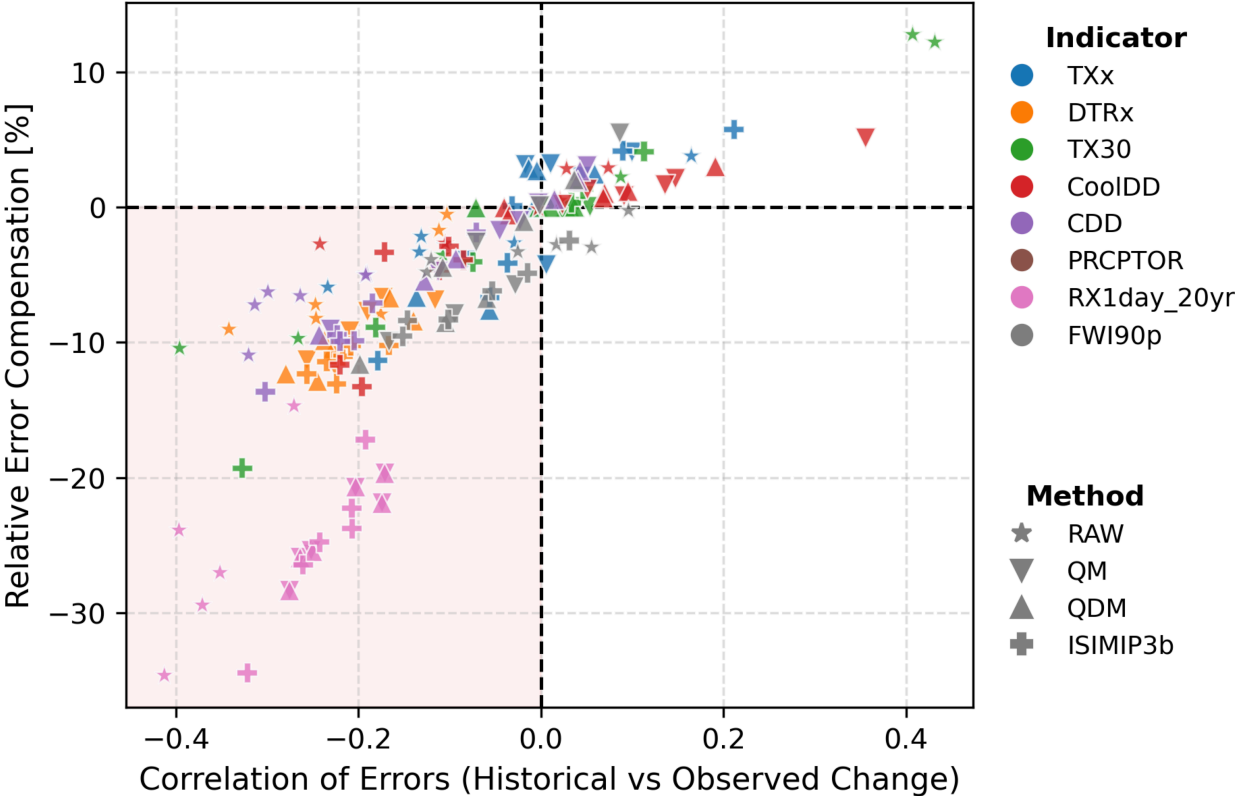


Figure 7. Scatter plot representing the spatial correlation between Historical and Observed Change errors (x-axis) versus the Relative Error Compensation [%] for the eight indicators (different colors), the raw climate models together with four different BA methods (distinct symbols) and the six climate models (repeated colors and symbols). Error compensation is estimated as the difference between the actual obtained error and the expected error estimated from the root sum of squared errors. Negative values indicate smaller than expected errors. Errors are relative to the expected error. Red area highlights where error compensation occurs due to the existence of systematic anticorrelation between Historical and Observed Change errors.



4.3 Added value of BA methods

When assessing the performance of post-processing techniques, such as bias adjustment, it is paramount to isolate and quantify the added value that each method brings compared to uncorrected climate model outputs. The NRMSESS effectively accounts for this, where values below or above zero indicate a degradation or improvement in performance, respectively, relative to raw model estimates. A complete elimination of baseline errors corresponds to a skill score of 1.0. In this context, Fig. 8a reinforces the aforementioned findings: while standard methods effectively reduce historical biases, they consistently fail to bring residual errors down to a negligible level. In fact, only three out of eight indicators (37.5%) exhibit an almost complete removal of initial historical bias ($>90\%$ error reduction), while an equal proportion of indicators achieve only a modest reduction ($\leq 50\%$). On average across indicators, standard methods leave approximately 30% of the original historical bias remaining in the corrected historical estimates, a notably high value given that observational data are explicitly used during training. Among standard methods, ISIMIP3B performs slightly worse (35% residual bias) than QM and QDM. This is exemplified by indicator RX1day_20yr, where the negative NRMSESS values in some models after ISIMIP3B adjustment indicate performance worse than that of the uncorrected simulations. The skill of standard BA methods in representing the baseline period also depends heavily on the underlying climate model, particularly for indicators such as TXx, DTRx, CDD, and RX1day_20yr, which exhibit an inter-model skill variability ranging from 25% to 60%. It is precisely here that IDS demonstrates one of its main advantages, it provides historical estimates fully anchored to observations independently of indicator and underlying climate model. A critical requirement for reliable downstream risk assessments that standard methods fail to guarantee.

The skill in representing observed trends (Fig. 8b) is ruled by the replication of most of the modeled changes by BA methods; thus, no added value is expected ($\text{NRMSESS} = 0$) for non-threshold-based indicators. For threshold-based indicators, the modification of the model trends adds some skill ($\sim 20\%$ bias reduction) only for the TX30 indicator across all methods. Interestingly, for CoolDD and CDD, corrected trends neither improve nor degrade the skill, meaning correct local updates are offset by incorrect updates. For PRCPTOT, however, only the direct application of modeled indicator-specific trends used by IDS, after adjusting the 1 mm/day dry-day threshold, adds 20% skill. In contrast, the distortion of projected trends in ISIMIP3B for this indicator leads to negative skill for most models, increasing errors by almost 25%, and exceeding 50% for one specific model. Conversely, in the case of multi-variable FWI90p, ISIMIP3B is the only method in which the little distortion of modeled trends by variable-specific parametric quantile mapping leads to a non-negligible skill gain (+5%). Another isolated but striking case is the negative skill, $\text{NRMSESS} < -0.5$, for one single model representing FWI90p observed trends after bias adjusted with QDM. All of this highlights once again the need for caution when using standard BA methods without prior verification.

Finally, evaluating skill scores for the future period (Fig. 8c) demonstrates that all BA methods deliver net added value relative to raw climate simulations ($\text{NRMSESS} > 0$). Skill for standard methods largely depends on their ability to reduce biases in the



historical period, and thus they exhibit a similar performance pattern for future estimates. Apart from the RX1day_20yr indicator, for which skill is close to 0 for standard methods, future NRMSESS spans a wide range, from 25 to 75% bias reduction, with large variability for indicator specific estimates within the model ensemble (up to 60%). IDS presents a much more consistent skill score (between 65 and 81% of future error reduction, one third of other methods' standard deviation) across these same indicators, with also much smaller inter-model variability (less than 25%). RX1day_20yr is also challenging for IDS, resulting in reduced skill for the future, but it still outperforms the other methods, with an NRMSESS of 0.2 vs 0.05 for QM/QDM. Averaging across all indicators, IDS clearly outperforms state of the art methods, with an NRMSESS of 0.68 compared to 0.54, representing a 25% improvement in skill or a 30% reduction in remaining error.

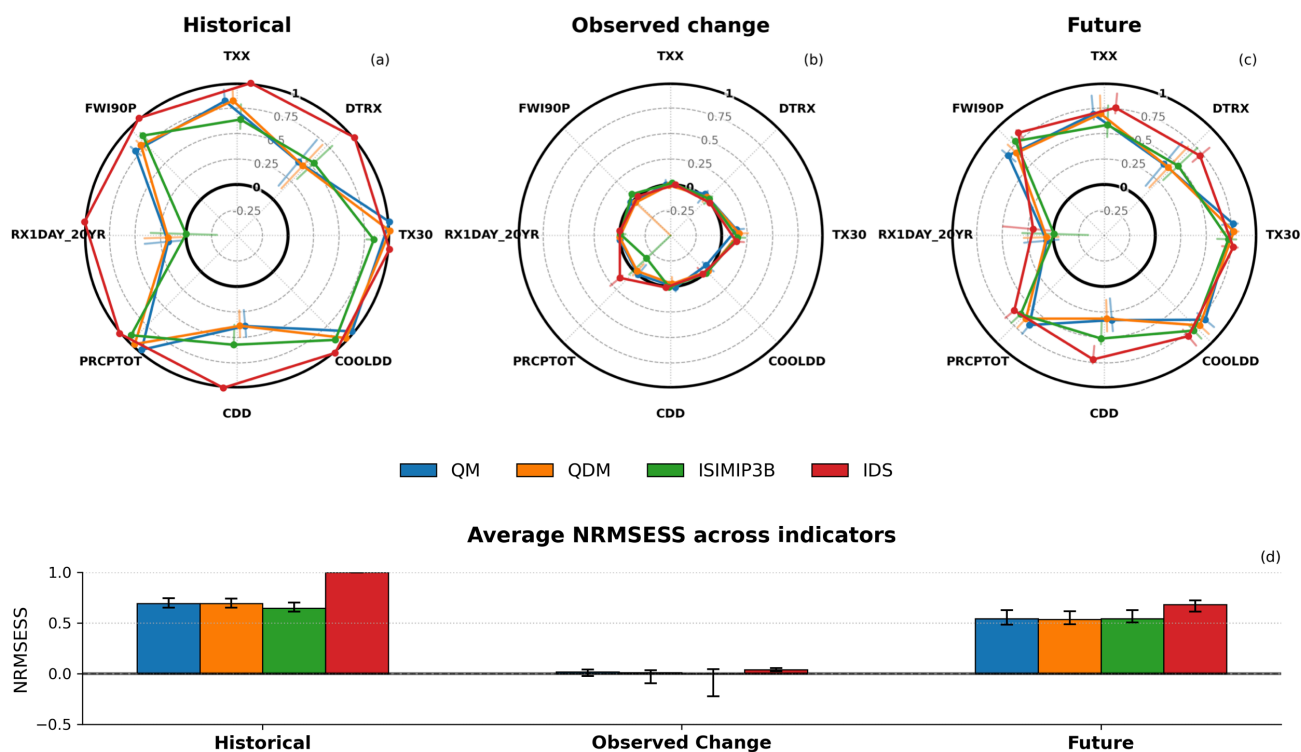


Figure 8. Top panel: Radar plots of the NRMSESS for each of the eight indicators are shown for the four BA methods (in color) considering the Historical (a), Observed Change (b) and Future (c) estimates. The circular lines represent the median of the multimodel ensemble and the vertical radial segments the minimum and maximum error within the ensemble. Bottom panel: Bar plots with the multimodel ensemble median NRMSESS averaged across all indicators for the three considered estimates and four different methods. Vertical segments represent the minimum and maximum error within the ensemble. Note that the larger the NRMSESS, the more accurate the method is.



4.4 Computational cost

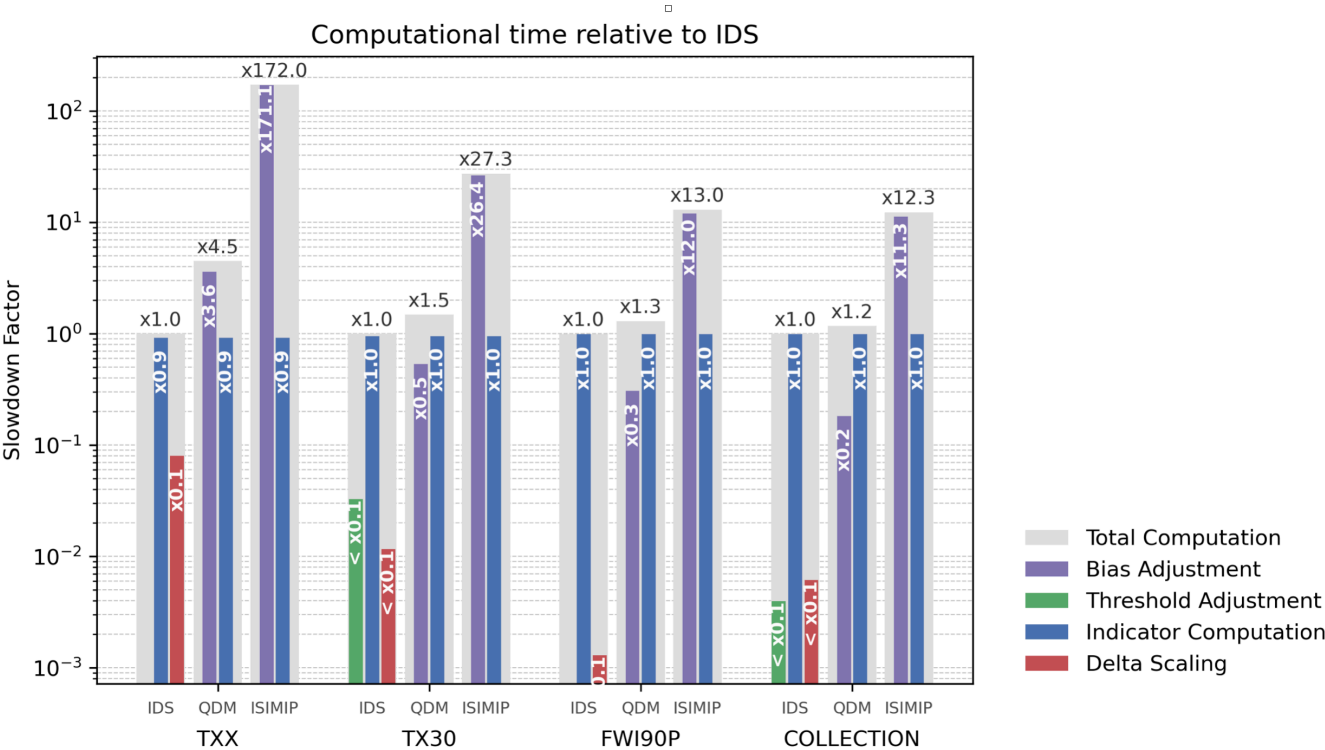
An additional benefit of the IDS approach is the reduction in computational cost, both in terms of resources and time. Rather than applying heavy BA procedures to the daily distribution of every input variable, IDS relies on light additional steps: it always requires the estimation of delta change factors and their application to the historical baseline, while some indicators might also require specific adjustments, such as the correction of indicator thresholds. This structural shift means that while IDS scales computationally only with the number and type of indicators considered, the resource requirements for standard methods also depend critically on the number of input variables to be adjusted. To better understand the computational burden for diverse types of indicators, we quantify the cost of these different steps in Fig. 9.

IDS presents the largest speedup factors (between 4.5x to 172x) compared to standard methods for simpler indicators, like TXx, where BA cost in standard methods represent several times the duration of the indicator's calculation itself. For threshold-based indicators like TX30, IDS requires a threshold adjustment, but this represents only 2% of the total processing time. For complex, computationally heavy indicators like FWI90p, the bias correction of input variables becomes secondary to the metric computation when using empirical quantile mapping methods like QDM. However, for the complex ISIMIP3B, the adjustment of input variables still represents by far the most significant part of the computation (13 times more than the indicator's calculation itself). Bypassing this BA step makes IDS 30% faster than QDM and consequently 13 times faster than ISMIP3B. Finally, when evaluating a collection of indicators, such as the eight analysed in this study where several bias-corrected input variables are reused across multiple indicators, IDS still maintains a clear advantage, reducing computational costs by 20% compared to QDM and by a factor of 12 compared to ISIMIP3B.

Another relevant consideration of the IDS approach is the flexibility it offers when the observational reference is updated with a different dataset. While standard methods require a computationally heavy re-adjustment of all climate models based on the new reference, along with the recalculation of all indicators, IDS avoids this bottleneck for non-threshold-based indicators. In this case, only the indicators corresponding to the new historical reference need to be computed. The bias-corrected future indicators are then efficiently updated by simply scaling this new baseline with the already established change factors. This same flexibility applies when updating climate indicators to consider new historical baselines or future periods. Once an indicator and its corresponding change factors are computed at a yearly or sub-annual scale covering the entire temporal extension, they can easily be aggregated into new timeframes at negligible cost. A common application of this efficiency occurs when working simultaneously with climate indicators represented in standard monotonic future periods and in Global Warming Levels (GWL).



568



569

570

571

572

573

574

575

576

Figure 9. Slowdown factors computed as the ratio between the total computational times for calculating four different types of indicators with three methods (QDM, ISMIP3B and IDS) relative to total IDS times. The collection case represents the calculation of the eight indicators considered in this study. Total computation times ratios are represented by grey bars while the diverse computational steps required are shown with different colors and are always relative to total IDS times. The QM method is not shown in this figure since it presents almost equal performance to QDM. To reduce the impact of the different computational efficiency of each method implementation, we only consider a single pixel and climate model for each climate indicator.

577

5 Conclusions

578

579

580

581

582

583

In the present study, a generalization of the Change Factor approach applied to diverse climate indicators is proposed and systematically evaluated to directly address the broad set of issues associated with current bias adjustment procedures repeatedly described in the literature (Maraun et al., 2017; Spuler et al., 2026). Moving the adjustment from the daily variable distributions to the indicator level, the proposed IDS represents a methodological shift toward a simpler, yet more robust and consistent solution based on three core principles that address key mechanisms underlying the issues identified in our evaluation: First, historical observations must be strictly preserved to guarantee the most realistic baseline upon which future



584 risk estimates are built. A future risk assessment grounded in an already flawed bias-adjusted historical baseline might lead to
585 inaccurate projections. Second, in the absence of an explicit physical or process-based justification for modifying the projected
586 change signal, future climate-change information should remain anchored to the change simulated by the underlying climate
587 model. Any deficiency in the change signal should remain the responsibility of the model itself, as purely statistical BA does
588 not aim to correct structural deficiencies in the underlying climate model. If an issue-specific correction is feasible, it should
589 fall under the expertise and responsibility of the corresponding practitioner, but never become the unintended artifact of a
590 generic unsupervised post-processing method. Third, to satisfy the first two principles simultaneously, adjustments must be
591 shifted from the underlying input data directly to the indicator computation itself. This use of indicator-specific adjustments
592 ensures the integrity of climate model data while deriving robust estimates from biased inputs. An example of this is the
593 adjustment of the corresponding indicator thresholds or the selection of a specific type of Delta Change Factor. These principles
594 allow IDS to directly avoid two critical and widespread problems: the misuse of BA to overcalibrate unrealistic climate models
595 representations (Chandel et al., 2024; Maraun et al., 2017) together with the distortion of observational baselines and projected
596 trends after BA (Casanueva et al., 2020; Dosio, 2016).

597 Beyond its conceptual benefits, IDS produces significantly more accurate estimates, systematically outperforming the three
598 benchmark methods across the evaluated indicators and models (QM, QDM, and ISIMIP3b) with a 23% smaller average error
599 (17% vs 22% NRMSE) and a 25% higher average skill (0.68 vs. 0.54 NRMSESS) for future period estimates. Furthermore,
600 IDS also provides markedly superior consistency, showing a performance variability across climate models and indicators
601 about three times smaller than other methods. A major contributor to this performance gap is the residual bias of the historical
602 baseline. Although complex standard methods calibrate the statistical historical distribution of each variable, they struggle to
603 actually replicate each specific climate indicator historical representation. Our results show standard methods leave behind,
604 after adjustment, roughly one third of the initial raw model historical bias with large variability across indicators, models and
605 methods. For instance, for indicator RX1day_20yr, the remaining historical bias after BA exceeds 75% for QM and QDM
606 methods, whereas it reaches 100% for ISIMIP3B, i.e. historical bias correction is negligible. Moreover, across most indicators,
607 locations with extreme outliers are present within the historical period with errors exceeding 50%, or even 100%, of the
608 corresponding indicator's global spatial variability, a severe usability problem for downstream risk assessments. In contrast,
609 IDS simply anchors to the observational baseline, ensuring systematic zero historical error independent of location, indicator
610 and climate model with respect to the observational reference.

611 The reproduction of model trends also reveals critical limitations in standard methods, in agreement with previous studies
612 (Casanueva et al., 2020; Spuler et al., 2024). Although all methods perform generally well preserving general trends for no
613 threshold-based indicators, they still fail to completely preserve model trends at local scale. The most striking representation
614 of this is the reversal of projected trends in several indicators, with standard methods presenting between 10 and 18% of
615 samples with a bias-adjusted projected change of the opposite sign compared to the raw simulation. Further analysis



616 demonstrates that these reversals generally provide no overall skill improvement, indicating they largely act as random
617 mathematical noise. A revealing exception occurs for the samples with largest reversals of the projected change, about ~1%
618 of all samples. While QDM lead to a strong degradation of model skill in representing trends for these samples, QM and the
619 parametric ISIMIP3B shows a significant gain in general but not for all models. The reduced number of samples for this
620 specific case, which contrasts with the neglectable improve in skill of general reversals of projected trends, avoids extracting
621 meaningful conclusions. Regarding the ability of standard BA to improve the representation of actual observed trends, their
622 trend-preserving nature makes they simply tend to inherit model errors adding no significant improvement. Yet, IDS adds
623 some minimal skill (5% average across indicators) due to the more precise correction of delta change factors on threshold-
624 based indicators (for PRCPTOT and TX30 indicators).

625 Therefore, the performance of standard BA methods in generating future estimates depends on two ostensibly independent
626 sources of error: the remaining biases in the historical estimate and the errors in reproducing the observed trends, which are
627 mostly driven by the skill of the climate models. However, our results show that these errors are not fully independent; rather,
628 they exhibit significant systematic correlation across specific indicators, models, and methods. This implies that historical and
629 trend errors can actually compensate for one another, or constructively compound, when combined to yield future estimates,
630 consequently leading to artificially smaller or larger errors than expected. Our dataset reveals a significant anticorrelation for
631 certain indicators, yielding up to an artificial 30% reduction in future estimate error. This introduces an additional source of
632 uncontrollable uncertainty in standard methods that, regardless of the underlying cause, leads to increased unreliability in
633 climate risk assessments. In contrast, in our evaluation and given an observational reference, the sole source of error in IDS
634 corresponds to the climate model's skill in projecting trends, ensuring that this uncertainty is fully identified, understood and
635 controlled.

636 Building on these results, IDS addresses the two main competing demands of the current “usability gap” of BA procedures
637 described by (Spuler et al., 2026): credibility and salience. Regarding scientific credibility, the fundamental design of IDS
638 provides a physically interpretable basis, straightforward interpretation, exact agreement with the observational baseline by
639 construction e, transparent uncertainty tracking, and simple standardization across diverse indicators. Furthermore, the
640 simplicity of IDS avoids the conventional assumption of stationarity of climate models’ biases, as these biases are entirely
641 bypassed in the delta approach. As a consequence, the remaining error in future estimates is more directly attributable to the
642 climate model's own deficiencies in projecting trends. This contrasts with standard methods, where violations of the stationarity
643 assumption can introduce additional, artificial and unexpected distortions to the projected trends. At the same time, IDS
644 achieves practical salience for users through its robustness across hazards, greater methodological transparency, and high
645 computational efficiency. Ultimately, by operating at the indicator level and avoiding the heavy processing of variable
646 distribution bias adjustment across decades of daily data, IDS provides massive computational gains, avoiding the x1.2 to x12



647 slowdown factors imposed by standard methods. This efficiency is complemented by accessible implementations fully
648 compatible with standard climate indicator libraries.

649 Despite these conceptual and empirical advantages, a primary limitation of IDS is its bounded scope: it is designed strictly to
650 scale climate indicators, not to generate the continuous, multivariate daily weather time series required for impact modeling.
651 In practice, however, climate indicators represent the vast majority of climate information processed in risk assessments and
652 decision-making. Furthermore, our results show that deriving future changes even for complex multivariate indicators, like
653 CoolDD and the highly non-linear FWI90p, from raw uncalibrated model inputs produced comparable or even lower future
654 errors than standard methods that meticulously calibrated the daily multivariate inputs first. This result can only be explained
655 by the fact that the compounding errors introduced throughout the standard BA pipeline exceed the error inherent in extracting
656 a change signal from uncalibrated inputs. In fact, initial tests showed that linearly pre-scaling raw models time series with the
657 historical observational reference to remove biases in models climatological mean values before indicator computation via
658 IDS did not lead to noticeable improvements in future error estimates, reinforcing the idea that biases in raw delta factors are
659 small. This suggests that a broader "impact delta" approach, where raw climate model outputs drive complex downstream
660 hydrological or agricultural impact models, with the resulting delta applied to observational impact baselines, may perform as
661 well as, or even outperform, the traditional pipeline of feeding bias-adjusted inputs into impact models, all while benefiting
662 from the aforementioned significant usability gains. Another important distinction and crucial requirement of the IDS approach
663 is the need for indicator-specific adjustments, which could theoretically pose challenges for certain highly concrete types of
664 indicators. While we cannot definitively rule out the existence of such cases, our extensive testing across dozens of diverse
665 indicators has not yet revealed any fundamental limitations where the IDS framework fails to provide a robust solution.

666 Finally, the versatility and transparency of the IDS approach suggest significant potential beyond climate projections. The
667 mathematical simplicity of extracting a pure change signal and scaling it against a robust baseline makes IDS a promising
668 candidate for seasonal-to-decadal climate predictions. In these forecasting contexts, correcting severe model drift and
669 accurately scaling forecasted anomalies to observed climatologies are critical challenges. By cleanly separating the physical
670 change signal from the baseline representation, this paradigm provides a transparent and operational pathway at these scales
671 as well. At climate projections scales, the results presented here demonstrate the practical applicability of the approach across
672 a diverse set of climate indicators. The framework has indeed been applied in the development of dedicated C3S climate
673 indicator datasets designed to address climate risk assessment requirements relevant to the EIB. These characteristics make
674 the framework relevant for large-scale indicators production, offering a practical and scalable solution for climate services
675 seeking to bridge the usability gap between complex climate modeling and robust, user-focused indicator data.



7 Code availability

A python-based implementation of the IDS framework used in this study is currently available at <https://github.com/Lobelia-Earth/xids>. Upon publication of the public preprint, we will make it accessible via Zenodo with a corresponding DOI.

8 Author contribution

JPI conceived the study, developed the methodology and corresponding code, performed all data analysis and curation, generated the figures, and wrote the manuscript. SH performed and wrote the computational efficiency analysis, contributed to the conceptualization and reviewed the manuscript and code. VEP, TR, LR and CC contributed to discussions and reviewed the manuscript.

9 Competing interests

"The authors declare that they have no conflict of interest."

10 Acknowledgements:

This research was supported by the European Centre for Medium-Range Weather Forecasts (ECMWF) under contract C3S2_431AtkinsRéalisis for the Copernicus Climate Change Service (C3S). With thanks to Sam Leader (Exeter UK), Francis Colledge (Exeter, UK) and Dr Steven Wade (Oxford, UK) of AtkinsRéalisis for review and technical discussion contributions in relation to the co-development of Climate-related Hazard indicators and relativity to climate risk assessment purposes through C3S2_431AtkinsRéalisis. The authors gratefully acknowledge the open-source Python community, specifically the core developers of xarray (Hoyer and Hamman, 2017) and xclim (Bourgault et al., 2023). During the preparation of this work, the authors used the Google Gemini AI tool to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

11 References

- Addor, N., Rohrer, M., Furrer, R., and Seibert, J.: Propagation of biases in climate models from the synoptic to the regional scale: Implications for bias adjustment, *JGR Atmospheres*, 121, 2075–2089, <https://doi.org/10.1002/2015JD024040>, 2016.
- Anandhi, A., Frei, A., Pierson, D. C., Schneiderman, E. M., Zion, M. S., Lounsbury, D., and Matonse, A. H.: Examination of change factor methodologies for climate change impact assessment, *Water Resources Research*, 47, 2010WR009104, <https://doi.org/10.1029/2010WR009104>, 2011.
- Arnbjerg-Nielsen, K.: Quantification of climate change effects on extreme precipitation used for high resolution hydrologic design, *Urban Water Journal*, 9, 57–65, <https://doi.org/10.1080/1573062X.2011.630091>, 2012.



- 703 Bourgault, P., Huard, D., Smith, T. J., Logan, T., Aoun, A., Lavoie, J., Dupuis, É., Rondeau-Genesse, G., Alegre, R., Barnes,
704 C., Laperrière, A. B., Biner, S., Caron, D., Ehbrecht, C., Fyke, J., Keel, T., Labonté, M.-P., Lierhammer, L., Low, J.-F., Quinn,
705 J., Roy, P., Squire, D., Stephens, A., Tanguy, M., and Whelan, C.: xclim: xarray-based climate data analytics, *JOSS*, 8, 5415,
706 <https://doi.org/10.21105/joss.05415>, 2023.
- 707 Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do
708 Methods Preserve Changes in Quantiles and Extremes?, *Journal of Climate*, 28, 6938–6959, [https://doi.org/10.1175/JCLI-D-](https://doi.org/10.1175/JCLI-D-14-00754.1)
709 14-00754.1, 2015.
- 710 Casanueva, A., Herrera, S., Iturbide, M., Lange, S., Jury, M., Dosio, A., Maraun, D., and Gutiérrez, J. M.: Testing bias
711 adjustment methods for regional climate change applications under observational uncertainty and resolution mismatch,
712 *Atmospheric Science Letters*, 21, e978, <https://doi.org/10.1002/asl.978>, 2020.
- 713 Chandel, V. S., Bhatia, U., Ganguly, A. R., and Ghosh, S.: State-of-the-art bias correction of climate models misrepresent
714 climate science and misinform adaptation, *Environ. Res. Lett.*, 19, 094052, <https://doi.org/10.1088/1748-9326/ad6d82>, 2024.
- 715 Deque, M.: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and
716 statistical correction according to observed values, *Global and Planetary Change*, 57, 16–26,
717 <https://doi.org/10.1016/j.gloplacha.2006.11.030>, 2007.
- 718 Dosio, A.: Projections of climate change indices of temperature and precipitation from an ensemble of bias-adjusted high-
719 resolution EURO-CORDEX regional climate models, *JGR Atmospheres*, 121, 5488–5511,
720 <https://doi.org/10.1002/2015JD024411>, 2016.
- 721 Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions “Should we apply bias correction to
722 global and regional climate model data?,” *Hydrol. Earth Syst. Sci.*, 16, 3391–3404, <https://doi.org/10.5194/hess-16-3391-2012>,
723 2012.
- 724 Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., and Zelinka, M.: Climate simulations: recognize the ‘hot
725 model’ problem, *Nature*, 605, 26–29, <https://doi.org/10.1038/d41586-022-01192-2>, 2022.
- 726 Hay, L. E., Wilby, R. L., and Leavesley, G. H.: A COMPARISON OF DELTA CHANGE AND DOWNSCALED GCM
727 SCENARIOS FOR THREE MOUNTAINOUS BASINS IN THE UNITED STATES¹, *J American Water Resour Assoc*, 36,
728 387–397, <https://doi.org/10.1111/j.1752-1688.2000.tb04276.x>, 2000.
- 729 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers,
730 D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara,
731 G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L.,
732 Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De Rosnay, P.,
733 Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, *Quart J Royal Meteor Soc*, 146, 1999–
734 2049, <https://doi.org/10.1002/qj.3803>, 2020.
- 735 Hoffmann, P., Menz, C., and Spekat, A.: Bias adjustment for threshold-based climate indicators, *Adv. Sci. Res.*, 15, 107–116,
736 <https://doi.org/10.5194/asr-15-107-2018>, 2018.
- 737 Hoyer, S. and Hamman, J.: xarray: N-D labeled Arrays and Datasets in Python, *JORS*, 5, 10, <https://doi.org/10.5334/jors.148>,
738 2017.



- 739 Iturbide, M., Casanueva, A., Bedia, J., Herrera, S., Milovac, J., and Gutiérrez, J. M.: On the need of bias adjustment for more
740 plausible climate change projections of extreme heat, *Atmospheric Science Letters*, 23, e1072,
741 <https://doi.org/10.1002/asl.1072>, 2022.
- 742 Lafon, T., Dadson, S., Buys, G., and Prudhomme, C.: Bias correction of daily precipitation simulated by a regional climate
743 model: a comparison of methods, *Intl Journal of Climatology*, 33, 1367–1381, <https://doi.org/10.1002/joc.3518>, 2013.
- 744 Lange, S.: Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0), *Geosci. Model Dev.*, 12,
745 3055–3070, <https://doi.org/10.5194/gmd-12-3055-2019>, 2019.
- 746 Maraun, D.: Bias Correcting Climate Change Simulations - a Critical Review, *Curr Clim Change Rep*, 2, 211–220,
747 <https://doi.org/10.1007/s40641-016-0050-x>, 2016.
- 748 Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M.
749 M., Hall, A., and Mearns, L. O.: Towards process-informed bias correction of climate change simulations, *Nature Clim*
750 *Change*, 7, 764–773, <https://doi.org/10.1038/nclimate3418>, 2017.
- 751 Olschewski, P., Laux, P., Wei, J., Böker, B., Tian, Z., Sun, L., and Kunstmann, H.: An ensemble-based assessment of bias
752 adjustment performance, changes in hydrometeorological predictors and compound extreme events in EAS-CORDEX,
753 *Weather and Climate Extremes*, 39, 100531, <https://doi.org/10.1016/j.wace.2022.100531>, 2023.
- 754 Olsson, J., Berggren, K., Olofsson, M., and Viklander, M.: Applying climate model precipitation scenarios for urban
755 hydrological assessment: A case study in Kalmar City, Sweden, *Atmospheric Research*, 92, 364–375,
756 <https://doi.org/10.1016/j.atmosres.2009.01.015>, 2009.
- 757 Padulano, R., Gomez-Mogollon, L. A., Napolitano, L., and Rianna, G.: Quantile-based bias-correction of extreme rainfall:
758 Pros & cons of popular methods for climate signal preservation, *Journal of Hydrology*, 653, 132814,
759 <https://doi.org/10.1016/j.jhydrol.2025.132814>, 2025.
- 760 Räisänen, J. and Rätty, O.: Projections of daily mean temperature variability in the future: cross-validation tests with
761 ENSEMBLES regional climate simulations, *Clim Dyn*, 41, 1553–1568, <https://doi.org/10.1007/s00382-012-1515-9>, 2013.
- 762 Spuler, F. R., Wessel, J. B., Comyn-Platt, E., Varndell, J., and Cagnazzo, C.: ibicus: a new open-source Python package and
763 comprehensive interface for statistical bias adjustment and evaluation in climate modelling (v1.0.1), *Geosci. Model Dev.*, 17,
764 1249–1269, <https://doi.org/10.5194/gmd-17-1249-2024>, 2024.
- 765 Spuler, F. R., Wessel, J. B., Jebeile, J., Zscheischler, J., and Shepherd, T. G.: Bias Adjustment and the Question of Usable
766 Climate Information: Methodological Assumptions and Value Judgments, *Bulletin of the American Meteorological Society*,
767 107, E79–E102, <https://doi.org/10.1175/BAMS-D-25-0022.1>, 2026.
- 768 Tootoonchi, F., Todorović, A., Grabs, T., and Teutschbein, C.: Uni- and multivariate bias adjustment of climate model
769 simulations in Nordic catchments: Effects on hydrological signatures relevant for water resources management in a changing
770 climate, *Journal of Hydrology*, 623, 129807, <https://doi.org/10.1016/j.jhydrol.2023.129807>, 2023.
- 771 Willems, P.: Revision of urban drainage design rules after assessment of climate change impacts on precipitation extremes at
772 Uccle, Belgium, *Journal of Hydrology*, 496, 166–177, <https://doi.org/10.1016/j.jhydrol.2013.05.037>, 2013.
- 773 Willems, P. and Vrac, M.: Statistical precipitation downscaling for small-scale hydrological impact investigations of climate
774 change, *Journal of Hydrology*, 402, 193–205, <https://doi.org/10.1016/j.jhydrol.2011.02.030>, 2011.