Environmental Data Science (2026), 5: e22, 1–12
doi:10.1017/eds.2026.10047

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Joint bias correction and downscaling of subseasonal forecasts via diffusion models

Maria Pyrina[1,2,3], Adel Imamovic[4], Dominik Bueeler[1,2,4], Christoph Spirig[4] and Daniela I.V. Domeisen[1,5]

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[2]Center for Climate Systems Modeling, C2SM, Switzerland
[3]European Centre for Medium-Range Weather Forecasts, ECMWF, Bonn
[4]Federal Office of Meteorology and Climatology, MeteoSwiss, Switzerland
[5]University of Lausanne, Switzerland
Corresponding author: Maria Pyrina; Email: maria.pyrina@env.ethz.ch

Received: 11 May 2026 UTC; **Revised:** 11 May 2026 UTC; **Accepted:** 04 June 2026 UTC

Keywords: complex orography; diffusion downscaling; subseasonal prediction; Switzerland; subseasonal diffusion downscaling

Abstract

Subseasonal forecasts provide valuable guidance for decision-making but remain limited by systematic biases and coarse spatial resolution, particularly over regions with complex orography. Quantile mapping (QM) is widely used for post-processing because it enables effective bias correction and downscaling, yet it applies a distributional correction and therefore neglects the case-specific dependence between forecasts and observations. To move toward a more integrated approach, we explore a diffusion-based generative framework for simultaneous probabilistic bias correction and high-resolution downscaling, with lead time explicitly included as a conditioning variable. We employ a conditional denoising diffusion model (DM) trained on high-resolution gridded observations (2 km) and conditioned on coarse-resolution ensemble-mean subseasonal forecasts (36 km) from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) subseasonal system. We focus here on a univariate temperature model and on training strategies. By learning conditional distributions of observations given forecasts and lead time, diffusion-based generative models enable case-dependent corrections that exploit forecast–observation relationships. Two training configurations are examined to assess how lead-time conditioning and data availability influence the learning of biases. Results show that the DM recovers realistic fine-scale temperature patterns and achieves meaningful lead-time-dependent bias correction across subseasonal lead times, even when trained on relatively limited datasets. However, compared to QM, the diffusion-based forecasts exhibit lower skill in probabilistic metrics such as the continuous ranked probability score (CRPS) and discrimination for extremes. Overall, this work represents a first step toward lead-time-aware subseasonal post-processing using DMs and highlights directions for future methodological development.

Impact Statement

Subseasonal forecasts play an increasingly important role in decision-making for sectors such as energy planning, agriculture, and public health. However, these forecasts are typically produced at coarse spatial

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

resolution and exhibit systematic biases, limiting their direct applicability for impact-oriented uses, particularly in regions with complex terrain. This study introduces a diffusion-based generative approach that performs lead-time-aware spatial downscaling and probabilistic bias correction of subseasonal temperature forecasts within a unified framework. Unlike conventional post-processing methods such as quantile mapping (QM), which correct forecasts based on marginal distributions and ignore case-specific forecast–observation relationships, the proposed method learns a conditional distribution of high-resolution temperature fields given the coarse-resolution forecast and lead time, enabling case-dependent probabilistic post-processing. This enables a computationally efficient and scalable pathway for operational implementation. The framework is demonstrated over Switzerland, where strong elevation gradients and local climate effects pose substantial challenges for forecasting systems. While traditional statistical methods remain competitive in probabilistic skill, this work establishes diffusion-based generative modeling as a viable foundation for next-generation subseasonal post-processing. By identifying current capabilities and remaining limitations, the study outlines clear directions for future development toward high-resolution subseasonal predictions that can better support decision-making.

1. Introduction

Subseasonal forecasting, typically defined as the prediction of atmospheric and surface variables from 2 weeks to 2 months ahead, fills a critical gap between medium-range weather prediction (up to two weeks ahead) and seasonal forecasts. This timescale is of high socioeconomic relevance for water resource management, wildfire and drought mitigation, agriculture, and disaster preparedness (Wilhite and Pulwarty 2017; White et al. 2022). Despite notable progress over the past decade (Pegion et al. 2019; Domeisen et al. 2022; Vitart et al. 2025), subseasonal forecasting remains a challenge (Merryfield et al. 2020).

Operational numerical weather prediction (NWP) systems have extended their forecast horizons into the subseasonal range using large global ensembles, typically operating at horizontal resolutions of 30–50 km (e.g., European Centre for Medium-Range Weather Forecasts [ECMWF] Integrated Forecasting System [IFS], National Aeronautics and Space Administration [NASA] Goddard Earth Observing System-Subseasonal to Seasonal [GEOS-S2S]). This resolution reflects a compromise between computational constraints and the need to represent large-scale coupled processes, such as the Madden–Julian oscillation (MJO), which are key sources of subseasonal predictability (Vitart 2017). However, systematic model biases in circulation, teleconnections, and surface variables grow with lead time and substantially degrade forecast skill across the subseasonal range (Pegion et al. 2019). In parallel, large-scale data-driven forecasting systems, including Pangu-Weather, GraphCast, and specialized subseasonal architectures such as FuXi-S2S, TianQuan-Climate, and Artificial Intelligence Forecasting System (AIFS)-S2S, have demonstrated improved representation of precipitation and teleconnections at lead times up to 45 days (Bi et al. 2023; Lang et al. 2026; Li et al. 2025). Nevertheless, like traditional NWP systems, these models remain constrained by relatively coarse spatial resolution.

Most subseasonal forecasting systems are designed to capture large-scale circulation patterns and, therefore, operate at resolutions insufficient to resolve mesoscale processes such as convection, diabatic heating, and flow–orography interactions, which occur at kilometer scales (Bauer et al. 2015). Errors arising from unresolved processes can propagate upscale, weakening large-scale teleconnections (e.g., MJO) and reducing forecast skill, particularly in the extratropics (Ham et al. 2019). These limitations are especially pronounced in regions of complex terrain, where sharp elevation gradients require kilometer-scale resolution for accurate representation (Lundquist et al. 2019). As a result, the lack of high-resolution, bias-corrected subseasonal forecasts remains a barrier for impact-oriented applications such as flood and drought prediction, renewable energy forecasting, and heat-stress assessment.

Deep generative models, and diffusion models (DMs) in particular, have emerged as powerful tools for probabilistic climate downscaling and bias correction (Xu et al., 2026; Mardani et al. 2025). Conditional DMs learn the distribution of fine-scale variability given coarse-scale forcing, enabling the generation of physically plausible high-resolution fields while explicitly representing uncertainty. Unlike deterministic post-processing approaches, DMs can implicitly correct systematic biases (Hatanaka et al. 2023) and

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

preserve multivariate dependencies when trained jointly (Lopez-Gomez et al. 2025), avoiding inconsistencies inherent in traditional sequential bias-correction and downscaling pipelines (Ehret et al. 2012; Maraun 2013).

Motivated by these advances, we develop a diffusion-based generative framework that simultaneously performs spatial downscaling and lead-time-dependent bias correction of ECMWF IFS subseasonal 2-meter temperature forecasts while generating a 50-member ensemble to represent forecast uncertainty. We evaluate the approach over Switzerland, a region characterized by pronounced orographic complexity spanning the Swiss Plateau and the Alpine mountain range. Strong elevation gradients and land–atmosphere interactions provide a stringent test for high-resolution subseasonal forecasting, particularly during summer, when convective processes dominate uncertainty. This setting enables a rigorous assessment of the ability of diffusion-based models to recover fine-scale structure and uncertainty from coarse, biased subseasonal forecasts.

2. Data and methods

2.1. Data and preprocessing

2.1.1. Observational data

The high-resolution target field used to train the DM is the 2 km gridded daily mean 2-meter temperature (2 t) over Switzerland, derived from the MeteoSwiss TabsD dataset (MeteoSwiss 2021). This observational product is generated through the spatial interpolation of approximately 90 homogenized long-term meteorological stations. The interpolation methodology incorporates a nonlinear estimation of vertical temperature profiles and the use of non-Euclidean distances, enabling a more accurate representation of temperature patterns influenced by cold-air pooling and downslope winds in mountainous valleys. The TabsD gridded record begins in 1961 and is continuously updated. At the time of data acquisition, data were available until 2020.

2.1.2. ECMWF subseasonal hindcasts

The coarse-resolution conditioning input for the DM consists of ECMWF subseasonal hindcasts. Subseasonal hindcasts are retrospective weather forecasts (re-forecasts) generated for past dates using the same numerical model and configuration as current real-time operations. ECMWF produces hindcasts to establish a robust *model climate*, which serves as a baseline for identifying significant anomalies, especially after the first prediction days. Hindcasts are used rather than real-time forecasts in order to maximize the size and consistency of the training dataset, which is essential for learning systematic model biases. To align with the availability of the observational dataset, we use ECMWF subseasonal hindcasts covering the period 2002–2021 from IFS cycle 47r3 (operational in 2022). This hindcast (and forecast) version is initialized twice weekly (Mondays and Thursdays) and extends up to 46 forecast days, with a native horizontal resolution of approximately 18 km for days 1–15, 36 km for days 16–46, and 137 vertical levels (see ECMWF Implementation of IFS Cycle 47r3 for details). The publicly available subseasonal hindcasts were distributed by ECMWF for download at the horizontal resolution of 36 km. Since cycle 47r3, several substantial upgrades have been implemented in the IFS, including the assimilation of 2-meter temperature observations, changes in the data assimilation systems, improvements to the land surface model, and a doubling of the number of ensemble members (Ingleby et al. 2024), leading to improvements in forecast skill (Roberts et al. 2024). Despite these advancements in the IFS, subseasonal forecasts still face persistent biases and resolution constraints, especially in topographically diverse regions. This study provides a framework designed to learn and correct these errors dynamically. As such, the approach remains highly relevant for current and future operational cycles.

2.1.3. Quantile-mapped subseasonal forecasts

The benchmark used for comparison in this study is the statistical downscaling framework employed operationally by MeteoSwiss for subseasonal forecasting. This framework utilizes empirical quantile mapping (QM) to bias correct and downscale ECMWF subseasonal 2 t ensemble forecasts against the

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

TabsD high-resolution observational field, following the approach validated by Monhart et al. (2018). Specifically, the technique maps the distribution of the raw model output onto a 2-km climate observational grid to produce localized forecasts.

2.1.4. Forecast verification metrics

To evaluate forecast performance, we employ a suite of standard meteorological verification metrics as detailed in Wilks (2011). The forecast mean error (ME) provides a primary assessment of deterministic accuracy and systematic error, indicating whether the ensemble mean consistently over- or underpredicts the observed values. To assess the quality of the probabilistic ensemble, we use the continuous ranked probability score (CRPS), which generalizes the mean absolute error (MAE) to probability distributions by integrating the squared difference between the forecast and observed cumulative distributions. Finally, the receiver operating characteristic (ROC) area under the curve (AUC) score is used to measure the system's discrimination ability, in our case specifically, its capacity to distinguish between the occurrence and nonoccurrence of temperature extremes above the observed 90th percentile. To further support the results of this work, additional verification diagnostics are provided in the Supplementary Material due to manuscript length constraints. These include the energy score (ES), which extends the CRPS to multivariate settings and evaluates the overall quality of ensemble forecasts, and rank histograms, which assess ensemble reliability by indicating whether the observations are statistically consistent with the forecast distribution.

2.2. Methodology

2.2.1. Generative diffusion framework

We employ a high-resolution meteorological downscaling model based on the conditional denoising diffusion probabilistic model (DDPM) framework (Mardani et al. 2025). The model is designed to generate $\mathbf{y}$, a high-resolution 2-meter temperature field, conditioned on $\mathbf{x}$, a low-resolution ensemble-mean input from the ECMWF subseasonal system, by learning the full conditional probability distribution $p(\mathbf{y}|\mathbf{x})$. During the diffusion process, structured noise is introduced, which, unlike standard independent Gaussian noise, accounts for the spatial correlations and multi-scale dependencies inherent in atmospheric fields. By iteratively removing this noise during inference, the model generates physically plausible fine-scale details, such as those driven by local topography, that are statistically consistent with high-resolution observations.

2.2.2. Network architecture and training

The core of our system is a U-Net backbone enhanced with the "elucidating the design space of diffusion-based generative models" (EDM) framework (Karras et al. 2022). We use EDM preconditioning to ensure training stability across varying noise scales. The input to the network is a two-channel tensor consisting of the noisy target image and the low-resolution condition $\mathbf{x}$, upscaled via bilinear interpolation and concatenated along the channel dimension. By using bilinearly upscaled forecasts as the conditioning input, we provide the DM with a spatially smooth but physically plausible starting point. In this configuration, the DM is explicitly tasked with a dual objective: first, to correct the systematic magnitude errors (bias) inherent in the low-resolution model data, and second, to recover the high-frequency spatial details (downscaling) absent in both the original low-resolution model and the subsequent interpolation step. The current verification evaluates the combined impact of these tasks, as the model is specifically architected to address both simultaneously. Additionally, we use lead time as a scalar auxiliary conditioning label to allow the model to learn lead-time-dependent bias patterns. Rather than predicting absolute temperature fields, the network is trained to model the residual between the bilinearly upscaled coarse input and the high-resolution reference. For numerical stability and faster convergence, this residual is standardized using a global Z-score normalization, where the mean and variance are computed across all grid points and lead times in the training set. The error scale is homogenized while preserving the underlying lead-time-dependent bias patterns and spatial structures. This allows the model to leverage

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

the auxiliary lead-time conditioning to learn and correct systematic errors without being hindered by the high absolute magnitude of the raw temperature values. The model is trained using a noise-weighted mean-square error (MSE) loss, which prioritizes the denoising performance at noise scales most critical for atmospheric feature reconstruction.

2.2.3. Probabilistic sampling and inference

To generate the final bias-corrected and downscaled product, we use a second-order deterministic sampler (Heun's method) with 40 steps. This allows us to transform Gaussian noise into high-resolution temperature fields that respect the large-scale constraints of the NWP input while adding realistic mesoscale variability. Starting from different random noise seeds for the same NWP input $\mathbf{x}$, we generate a 50-member ensemble. This ensemble is used to quantify the forecast uncertainty and provide probabilistic bias correction, effectively mapping the biased, coarse-grained dynamical model output to the observational data.

2.2.4. Training strategy

The ECMWF hindcast dataset comprises 520 hindcasts, each consisting of 11 ensemble members (one control and ten perturbed) initialized over 20 summers (June, July, August) from 2002 to 2021. We adopt a stratified training approach, in which the DM is trained separately for each model start date (i.e., initialization date). This strategy is chosen to account for the strong subseasonal non-stationarity of the atmospheric state; by isolating start dates, the model can more effectively learn the specific seasonal evolution of biases and land–atmosphere feedbacks unique to each phase of the summer. While a common alternative would be to pool all start dates and utilize a seasonal conditioning label, we chose separate training for each start date to maximize the model's sensitivity to date-specific atmospheric regimes. This strategy avoids the risk where the model might average the distinct physical biases of early-summer snowmelt periods with late-summer dry regimes (Orth and Seneviratne 2015). For each start date, we train for 3000 epochs using the ensemble-mean 2 t fields from 2002 to 2017 without incorporating additional climate variables. Model selection and hyperparameter tuning are performed using the 2018–2019 validation period, while the year 2020 is reserved as an independent test set. Moreover, note that even though the target consists of daily 2 t fields, the evaluation on the test period is aggregated to a weekly scale in order to enable a consistent assessment of subseasonal forecast skill.

Motivated by the lead-time-dependent biases reported for subseasonal forecasts over Switzerland (i.e., Figure 12 in Monhart et al. [2018]), we evaluate two distinct training strategies. The first, an all-lead-time approach, utilizes the ensemble-mean fields across the full 46-day forecast horizon, incorporating daily lead-time information as a conditioning variable during training. The second, a week 1–restricted approach, excludes lead-time information and limits the training data strictly to lead week 1. To compensate for the significantly reduced sample size in the week 1–restricted approach, we increase the training duration to 6000 epochs to ensure model convergence. For the all-lead-time approach, we train for 24 hours using four graphics processing units (GPUs), while for the week 1–restricted approach we train for 48 hours. The training was performed on Nvidia Grace Hopper processors in MeteoSwiss' Research and Development Cluster Balfrin, at the Swiss National Supercomputing Centre (CSCS).

3. Results and discussion

Both training strategies demonstrate a similar capability to downscale the ECMWF forecasts, producing spatially detailed temperature fields that exhibit patterns similar to those observed in the reference dataset (Figure 1). In terms of bias correction (Figure 2), complete bias removal remains challenging for all methods at these forecast ranges, particularly given the limited training data and the inherent uncertainty at longer lead times. Both DM approaches provide notable improvements of ME over the interpolated ECMWF data, although the improvements are somewhat smaller, though still comparable, to those achieved by QM. For lead weeks 1 and 2, the all-lead-time strategy (DM_all) outperforms the week

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

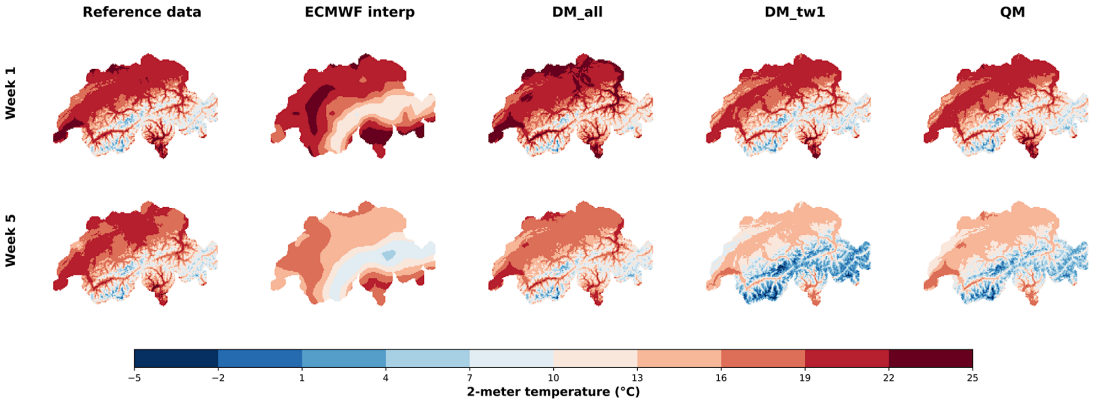


Figure 1. *Weekly 2 t from example forecasts initialized on 15 August 2020 (columns 2–5), together with the reference observational dataset (column 1). The panels display forecasts from the interpolated ECMWF IFS cycle 47r3 (second column), the DM_all approach (third column), the DM_tw1 approach (fourth column), and the QM approach (fifth column). The results are given for lead week 1 (first row) and lead week 5 (second row).*

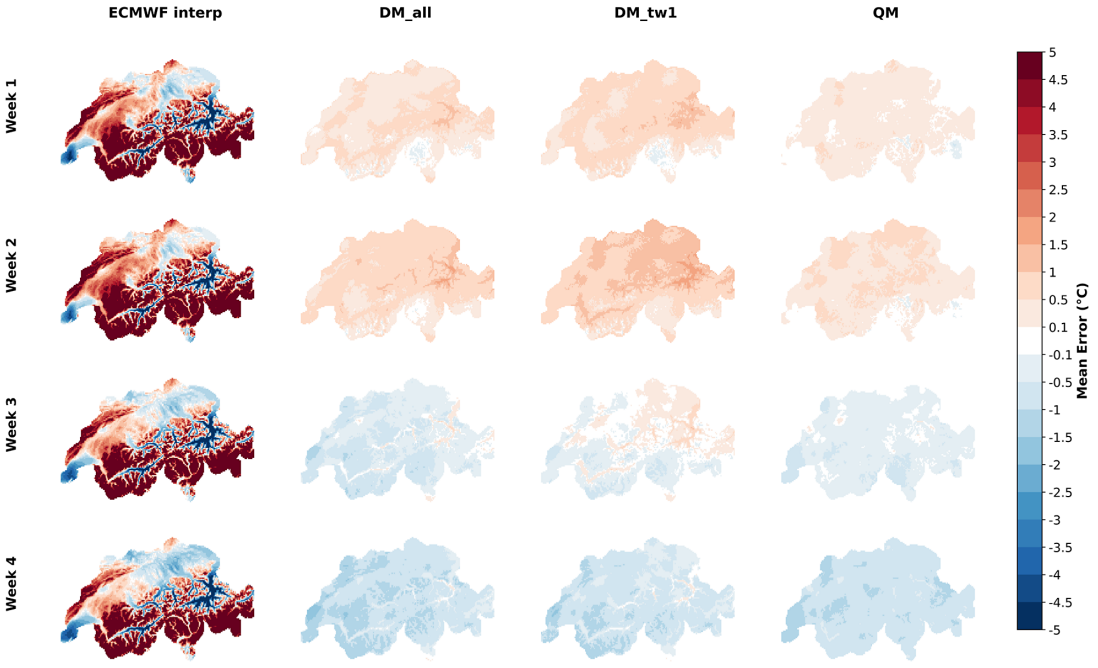


Figure 2. *Weekly 2 t ME derived from all model initializations between June and August 2020 for the interpolated ECMWF IFS cycle 47r3 (first column), the DM_all approach (second column), the DM_tw1 approach (third column), and the QM approach (fourth column). ME values closer to zero indicate a better score.*

1–restricted approach (DM_tw1) over most of Switzerland, as indicated by smaller values of ME over most areas (Figure 2). In particular, for week 1 predictions, the all-lead-time approach reduces the ME by roughly twice as much as the week 1–restricted strategy, suggesting that the DM effectively exploits information from the daily week 1 lead times. For weeks 3 and 4, as well as for longer lead times not

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

shown here (weeks 5 and 6), the week 1–restricted approach achieves a bias correction that is comparable to or slightly better than that of the all-lead-time strategy. However, as the differences between the two DM approaches are small for longer lead times, a larger sample size would be required to robustly assess their relative bias-correction performance. The similar bias correction observed for longer lead times between the two training strategies is partly attributable to the relative stability of weekly mean biases in the interpolated ECMWF forecasts at lead times longer than 2 weeks and to the small bias differences across the different lead weeks for several Swiss locations (see Supplementary Figure S1). Specifically, the bias can remain relatively stable either across all lead times (e.g., Zurich and Geneva in Supplementary Figure S1) or between shorter (first two weeks) and longer lead times (e.g., Locarno and Sion in Supplementary Figure S1). Although the DM_tw1 approach is trained on a smaller dataset than DM_all, it benefits from longer training time and is still exposed, through week 1 data, to bias magnitudes comparable to those at longer lead times. Moreover, the DM learns the overall data distribution rather than performing simple location-to-location adjustments, which likely contributes to its ability to generalize across lead times. It is also important to emphasize that the inputs to the DM consist of ECMWF forecasts interpolated from 36 km to 2 km resolution. This interpolation step can amplify the raw bias at all lead times, particularly over orographically complex regions. Moreover, the ECMWF model outputs are compared against observational data that were not seen during training. Consequently, discrepancies arising from resolution differences and the use of independent observational references may have a larger impact on model errors than the lead-time-dependent drift itself. The observed bias stability at specific locations and/or lead times is consistent with the results of Dutra et al. (2021) (see their Figure 3a,b) that evaluate subseasonal ECMWF forecasts against fifth generation ECMWF atmospheric reanalysis (ERA5), the dataset used for the ECMWF model initialization (Vitart et al. 2019; Hersbach et al. 2020).

The CRPS of the raw ECMWF forecasts is substantially reduced by both DM approaches, particularly over orographically complex regions, although the reduction is generally smaller than that achieved by QM (Figure 3). For lead week 1, the QM method reduces the raw ECMWF CRPS by nearly 1 °C over the Swiss Plateau and by more than 5 °C in complex terrain. The DM approaches reduce the ECMWF CRPS by approximately 0.5 °C over the Swiss Plateau and by about 4 °C over orographically complex regions, where the raw ECMWF CRPS reaches about 1 °C and exceeds 5 °C, for these regions, respectively. While the all-lead-time DM approach yields marginally lower CRPS improvements compared to the week 1–restricted strategy, the magnitude of the differences is small to support a statistically robust distinction between the two methods.

Finally, we assess the ability of the different approaches to predict hot extremes, defined as temperatures exceeding the observed 90th percentile (Figure 4). Discrimination skill is evaluated using the ROC/AUC score, where values of 0.5 or lower indicate no skill beyond random chance. The ROC/AUC metric accounts for both the hit rate and the false alarm rate across all possible decision thresholds, making it well suited for evaluating probabilistic predictions of extremes. Across all lead weeks, the QM approach attains AUC scores exceeding 0.7 in most of Switzerland. Although slightly lower, both DM strategies also demonstrate skill, with regional AUC values consistently above 0.5 for all lead weeks and both approaches, except for the all-lead-time strategy at lead week 3. From lead weeks 1 to 4, the week 1–restricted DM strategy generally outperforms the all-lead-time approach, whereas for lead weeks 5 and 6 (not shown), the all-lead-time strategy exhibits better discrimination. Across all methods, the ROC score generally decreases with increasing lead time from 1 to 6 weeks; however, both DM strategies display an unusually large reduction at lead week 3. A similar anomaly at lead week 3 is also evident in the CRPS (Figure 3) and may be related to the change in forecast resolution at this lead time in the ECMWF system (see Methods). For lead week 1, QM achieves the highest AUC, exceeding 0.95 for nearly all Swiss regions, followed by the DM approaches with regional values mostly above 0.9. This indicates that while QM performs best overall, both DM methods retain substantial skill in discriminating hot extremes across lead times.

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

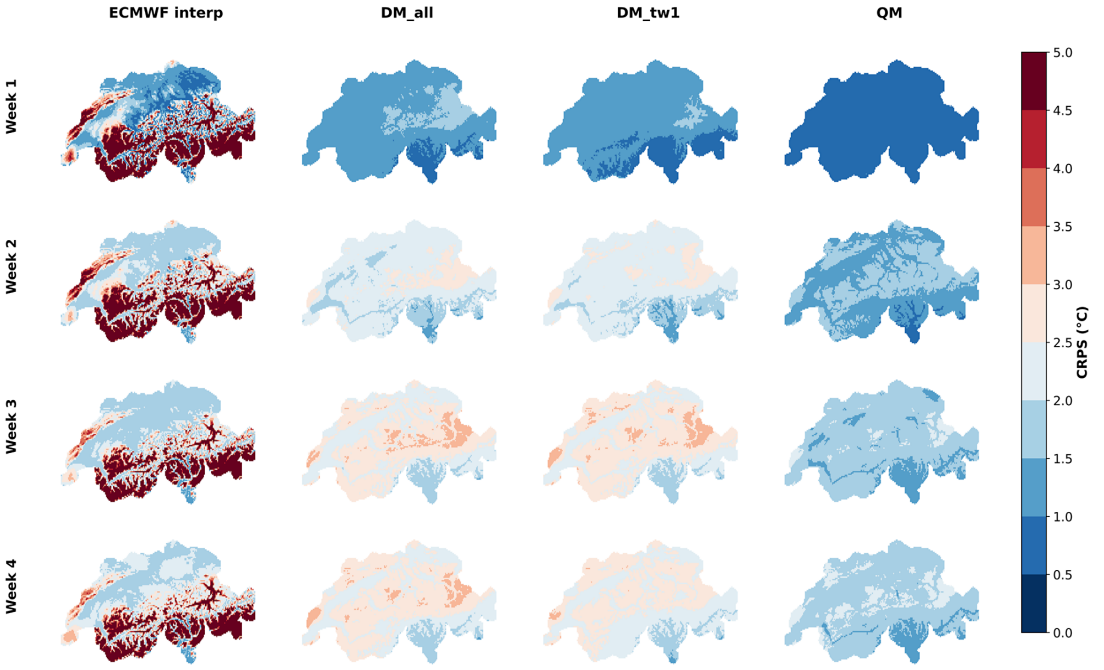


Figure 3. As Figure 2, but for the weekly 2 t CRPS. CRPS values closer to zero indicate a better score.

Although DM_all is exposed to a larger training set, the inclusion of longer lead times (weeks 2–6), which are characterized by increased ensemble spread, results in a smoother ensemble mean used as training input. The competitive performance of DM_tw1 for longer lead times suggests that, despite being trained on a smaller dataset, it effectively leverages the higher signal-to-noise ratio of week 1 data. This indicates that the DM is able to learn the fundamental relationship between large-scale atmospheric conditions and fine-scale topographic features primarily from these high-quality inputs. Importantly, this learned relationship appears to generalize well to later lead times, even though the model is not explicitly trained on them.

4. Conclusion

This study introduces a generative diffusion-based framework for the simultaneous downscaling and bias correction of subseasonal temperature forecasts over complex terrain. By conditioning a probabilistic DM on coarse-resolution ECMWF subseasonal forecasts and explicitly incorporating lead-time information, we bias correct and downscale coarse horizontal resolution (36 km) ECMWF subseasonal hindcasts into high horizontal resolution (2 km) daily mean temperature fields over Switzerland. The proposed approach generates physically plausible, high-resolution temperature fields while simultaneously accounting for forecast uncertainty through ensemble sampling. Both of these aspects are important for impact-oriented applications.

We have evaluated two distinct training strategies on a univariate 2 t model. Specifically, the all-lead-time approach reduces the ME in week 1 predictions by approximately 50% more compared to the week 1–restricted strategy. For longer lead times, both methods show comparable performance, likely due to the relative stability of weekly mean biases in the underlying coarse-resolution forecasts as compared to the reference dataset that was never seen by the model and is based on spatially interpolated observations. While the baseline QM method retains an advantage in terms of CRPS and extreme event discrimination (ROC/AUC), the DMs successfully reduce raw ECMWF CRPS by up to 4 °C in orographically complex

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

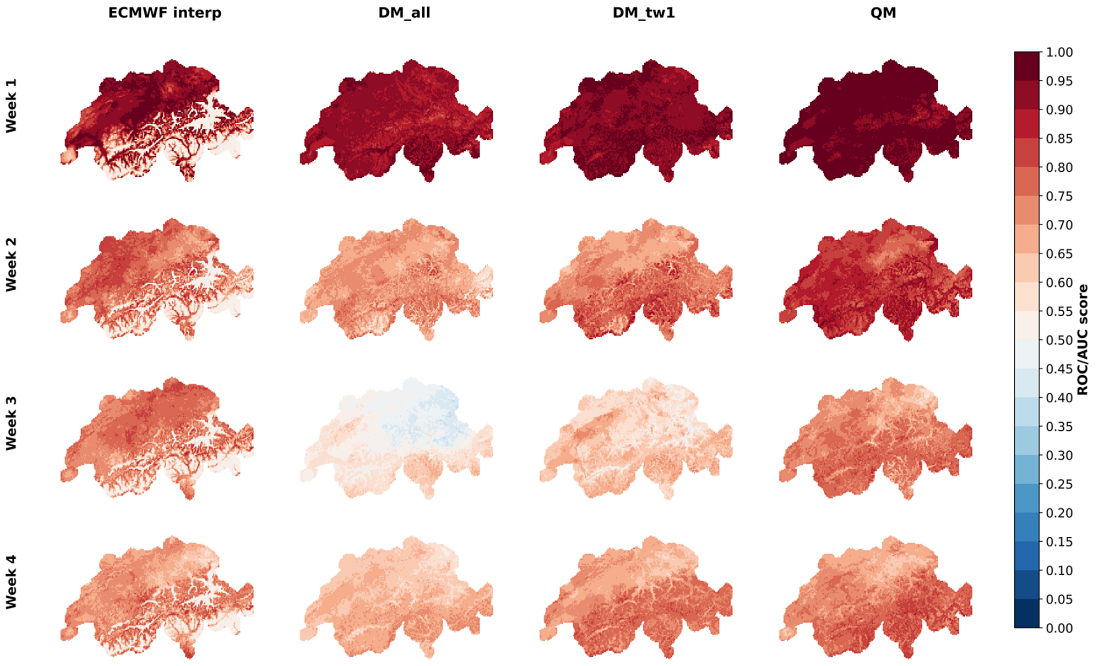


Figure 4. As Figure 2, but for the weekly ROC/AUC score of 2 t extremes above the observed 90th percentile. ROC/AUC values closer to 1 indicate a better score, and values equal to or lower than 0.5 indicate no skill beyond random chance in discriminating between extreme and non-extreme events.

regions. QM's strong performance reflects its targeted optimization of marginal distributions, making it a difficult baseline to outperform, particularly in probabilistic verification metrics.

A key contribution of this work is the explicit examination of lead-time conditioning and training strategies in generative downscaling for the subseasonal range. By contrasting an all-lead-time training strategy with a week 1–restricted approach, we show that DMs can flexibly correct bias structures across lead times, while remaining robust even when trained on relatively limited datasets. Importantly, our results indicate that effective subseasonal post-processing does not necessarily require extensive training data across all lead times, making the approach attractive for applications where data availability or computational resources are constrained. Beyond forecast skill, the proposed framework offers advantages for operational use. Unlike traditional QM, the DM generates its own ensemble from a single forecast input, reducing storage needs and enabling relatively fast inference (50 members in 40 minutes on a single GPU). In addition, its generative design allows new predictors or variables to be easily incorporated, facilitating extension to other surface variables or challenging regimes such as wintertime temperature inversions.

While this study focuses on summer temperature forecasts over Switzerland, chosen as a challenging test case due to strong orographic effects and pronounced summer biases, the methodological insights are broadly transferable. The proposed diffusion-based framework is not tied to a specific model cycle, geographic region, or atmospheric variable and is therefore expected to remain applicable as dynamical forecasting systems continue to evolve. In this work, the DM is conditioned on the ensemble-mean forecast in order to isolate and systematically evaluate the impact of the two training strategies on bias-correction performance. This design choice, however, may limit the generalization of the model and its ability to fully represent ensemble spread.

Several promising avenues for future development, therefore, emerge from this study. Training the DM on all ensemble members or on a single forecast (e.g., control forecast) could improve the representation of forecast uncertainty and extremes and potentially enhance probabilistic skill. In addition, jointly

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

training across multiple initialization dates may improve generalization across atmospheric regimes while retaining lead-time awareness. Further extensions could include multivariate downscaling, incorporation of additional physical predictors, or the prediction of impact-relevant indices. Although these developments are beyond the scope of the present work, they represent natural and well-motivated next steps toward fully exploiting diffusion-based generative models for subseasonal forecasting.

Overall, this work provides a concrete and operationally relevant benchmark for applying lead-time-informed diffusion-based generative models to subseasonal forecasting. It demonstrates that such models can effectively bridge the gap between coarse-resolution dynamical forecasts and the high-resolution probabilistic information required for impact-oriented applications, thereby contributing a practical tool to the growing field of climate informatics.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2026.10047.

Data availability statement. ECMWF IFS hindcast data are available from the ECMWF public data archive at https://apps.ecmwf.int/archive-catalogue/. The observational TabsD dataset can be accessed through MeteoSwiss' Open Government Data infrastructure (https://www.meteoschweiz.admin.ch/service-und-publikationen/service/open-data.html). The DM code is publicly available at https://github.com/marpyr/Diffusion_Downscaling.

Acknowledgments. We acknowledge financial support from the Swiss drought program (https://www.meteoswiss.admin.ch/about-us/research-and-cooperation/projects/2024/drought-program.html). We extend our gratitude to MeteoSwiss for the access to the TabsD observational products, as well as the subseasonal forecast archive. M.P. and D.B. acknowledge financial support from the Collaborative Research on Science and Society (CROSS) program of École Polytechnique Fédérale de Lausanne (EPFL) and Université de Lausanne (UNIL). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 847456) and from a grant from the Swiss National Supercomputing Centre (CSCS) under project ID s1144. We would also like to thank and acknowledge Joffrey Dumont Le Brazidec (ECMWF), Maxim Samarin (Swiss Data Science Center), Jakob Schloer (ECMWF), and Steffen Tietsche (ECMWF) for fruitful discussions.

Author contributions. M.P. conceptualized and conducted the study, built the DM, analyzed and visualized the data, and wrote the original manuscript draft. A.I. provided the hindcast data and quantile-mapped forecasts and conceptualized the study. D.B. conceptualized the study. C.S. and D.D. acquired funding, conceptualized the study, and supervised the study. All authors reviewed the manuscript.

Competing interests. The authors confirm that they have no competing interests.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Provenance statement. This article is part of the Climate Informatics 2026 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer-review process.

References

Bauer P, **Thorpe A and Brunet G** (2015) The quiet revolution of numerical weather prediction. *Nature 525*(7567), 47–55. https://doi.org/10.1038/nature14956.

Bi K, **Xie L**, **Zhang H**, **Chen X**, **Gu Q and Tian Q** (2023) Accurate medium-range global weather forecasting with 3D neural networks. *Nature 619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3.

Chen, L, **Zhong, X**, **Zhang, F**, **Cheng, Y**, **Xu, Y**, **Qi, Y**, & **Li, H** (2023). FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *NPJ Climate and Atmospheric Science*, *6*(1), 190.

Domeisen DI, **White CJ**, **Afargan-Gerstman H**, **Muñoz ÁG**, **Janiga MA**, **Vitart F and Tian D** (2022) Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe. *Bulletin of the American Meteorological Society 103*(6), E1473–E1501. https://doi.org/10.1175/BAMS-D-20-0221.1.

Dutra E, **Johannsen F and Magnusson L** (2021) Late spring and summer subseasonal forecasts in the northern hemisphere midlatitudes: Biases and skill in the ECMWF model. *Monthly Weather Review 149*(8), 2659–2671.

Ehret U, **Zehe E**, **Wulfmeyer V**, **Warrach-Sagi K and Liebert J** (2012) Hess opinions "should we apply bias correction to global and regional climate model data?". *Hydrology and Earth System Sciences 16*(9), 3391–3404. https://doi.org/10.5194/hess-16-3391-2012.

Ham Y-G, **Kim J-H and Luo J-J** (2019) Deep learning for multi-year ENSO forecasts. *Nature 573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7.

Hatanaka Y, **Glaser Y**, **Galgon G**, **Torri G and Sadowski P** (2023) Diffusion models for high-resolution solar forecasts. *arXiv preprint* arXiv:2302.00170.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **De Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **de Rosnay P**, **Rozum I**, **Vamborg F**, **Villaume S and Thépaut J-N** (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049.

Ingleby B, **Arduini G**, **Balsamo G**, **Boussetta S**, **Ochi K**, **Pinnington E and de Rosnay P** (2024) Improved two-metre temperature forecasts in the 2024 upgrade. *ECMWF Newsletter 178*, 24–29.

Karras T, **Aittala M**, **Aila T and Laine S** (2022) Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, *35*, 26565–26577.

Lang S, **Alexe M**, **Clare MCA**, **Roberts C**, **Adewoyin R**, **Bouallègue ZB**, **Chantry M**, **Dramsch J**, **Dueben PD**, **Hahner S**, **Maciel P**, **Prieto-Nemesio A**, **O'Brien C**, **Pinault F**, **Polster J**, **Raoult B**, **Tietsche S and Leutbecher M** (2026) Aifs-crps: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *NPJ Artificial Intelligence 2*(1), 18.

Li G, **Liu X**, **Cao S**, **Liang H**, **Chen M**, **Zhang L**, **Zhang J**, **Wang J**, **Jin M**, **Zheng J and Fu H** (2025) TianQuan-climate: A subseasonal-to-seasonal global weather model via incorporate climatology state. arXiv e-prints, arXiv-2504.

Lopez-Gomez I, **Wan ZY**, **Zepeda-Núñez L**, **Schneider T**, **Anderson J and Sha F** (2025) Dynamical-generative downscaling of climate model ensembles. *Proceedings of the National Academy of Sciences 122*(17), e2420288122. https://doi.org/10.1073/pnas.2420288122.

Lundquist J, **Abel MR**, **Gutmann E and Kapnick S** (2019) Our skill in Modeling Mountain rain and snow is bypassing the skill of our observational networks. *Bulletin of the American Meteorological Society 100*(12), 2473–2490. https://doi.org/10.1175/BAMS-D-19-0001.1.

Maraun D (2013) Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate 26*(6), 2137–2143. Supports the argument on the limitations of sequential post-processing. https://doi.org/10.1175/JCLI-D-12-00821.1.

Mardani M, **Brenowitz N**, **Cohen Y**, **Pathak J**, **Chen C.Y**, **Liu CC**, **Vahdat A**, **Nabian MA**, **Ge T**, **Subramaniam A**, **Kashinath K**, **Kautz J and Pritchard M** (2025) Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment 6*(1), 124.

Merryfield WJ, **Baehr J**, **Batté L**, **Becker EJ**, **Butler AH**, **Coelho CAS**, **Danabasoglu G**, **Dirmeyer PA**, **Doblas-Reyes FJ**, **Domeisen DIV**, **Ferranti L**, **Ilynia T**, **Kumar A**, **Müller WA**, **Rixen M**, **Robertson AW**, **Smith DM**, **Takaya Y**, **Tuma M**, **Vitart F**, **White CJ**, **Alvarez MS**, **Ardilouze C**, **Attard H**, **Baggett C**, **Balmaseda MA**, **Beraki AF**, **Bhattacharjee PS**, **Bilbao R**, **de Andrade FM**, **DeFlorio MJ**, **Díaz LB**, **Ehsan MA**, **Fragkoulidis G**, **Gonzalez AO**, **Grainger S**, **Green BW**, **Hell MC**, **Infanti JM**, **Isensee K**, **Kataoka T**, **Kirtman BP**, **Klingaman NP**, **Lee JY**, **Mayer K**, **McKay R**, **Mecking JV**, **Miller DE**, **Neddermann N**, **Justin Ng CH**, **Ossó A**, **Pankatz K**, **Peatman S**, **Pegion K**, **Perlwitz J**, **Recalde-Coronel GC**, **Reintges A**, **Renkl C**, **Solaraju-Murali B**, **Spring A**, **Stan C**, **Sun YQ**, **Tozer CR**, **Vigaud N**, **Woolnough S and Yeager S** (2020) Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society 101*(6), E869–E896. https://doi.org/10.1175/BAMS-D-19-0037.1.

MeteoSwiss (2021) Meteoswiss spatial climate analyses: Documentation of datasets for users. Available at https://www.meteoswiss.admin.ch/climate/the-climate-of-switzerland/spatial-climate-analyses.html.

Monhart S, **Spirig C**, **Bhend J**, **Bogner K**, **Schär C and Liniger MA** (2018) Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres 123*(15), 7999–8016. https://doi.org/10.1029/2017JD027923.

Orth R and Seneviratne SI (2015) Impact of soil moisture initialisation on Swiss summer daily maximum temperature forecasting. *Agricultural and Forest Meteorology 206*, 19–27.

Pegion K, **Kirtman B**, **Becker E**, **Collins D**, **LaJoie E**, **Burgman R**, **Bell M**, **DelSole T**, **Min D-H**, **Zhu Y**, **Li W**, **Sinsky E**, **Guan H**, **Gottschalck J**, **Metzger EJ**, **Barton NP**, **Achuthavarier D**, **Marshak J**, **Koster RD**, **Lin H**, **Gagnon N**, **Bell M**, **Tippett MK**, **Robertson AW**, **Sun S**, **Benjamin SG**, **Green BW**, **Bleck R and Kim H** (2019) The subseasonal experiment (subx): A multimodel subseasonal prediction experiment. *Bulletin of the American Meteorological Society 100*(10), 2043–2060. https://doi.org/10.1175/BAMS-D-18-0270.1.

Roberts CD, **Ingleby B**, **Geer A**, **Hólm E**, **Janousek M**, **Prates F and Rodwell M** (2024) Ifs upgrade improves near-surface wind and temperature forecasts. *ECMWF Newsletter 181*, 16–25.

Xu W, **Lu Y**, **Wang S and Liu TR** (2026) Improving data fidelity via diffusion model-based correction and super-resolution. *Journal of Computational Physics*, 114883.

Vitart F (2017) Madden—Julian oscillation prediction and teleconnections in the s2s database. *Quarterly Journal of the Royal Meteorological Society 143*(706), 2210–2220. https://doi.org/10.1002/qj.3079.

Vitart F, **Balsamo G**, **Bidlot J**, **Lang S**, **Tsonevsky I**, **Richardson D and Alonso Balmaseda M** (2019) Use of era5 to Initialize Ensemble Re-Forecasts. ECMWF Technical Memorandum. Available at https://www.ecmwf.int/en/elibrary/80956-use-era5-initialize-ensemble-re-forecasts.

Vitart F, **Robertson A**, **Brookshaw A**, **Caltabiano N**, **Coelho C**, **de Coning E**, **Dirmeyer P**, **Domeisen D**, **Hirons L**, **Kim H**, **Lin H**, **Kumar A**, **Molod A**, **Robbins J**, **Segele Z**, **Spillman CM**, **Stan C**, **Takaya Y**, **Woolnough S**, **White CJ and Wu T** (2025)

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press

The WWRP/WCRP s2s project and its achievements. *Bulletin of the American Meteorological Society 106*(5), E791–E808. https://doi.org/10.1175/BAMS-D-24-0047.1.

White CJ, **Domeisen DI**, **Acharya N**, **Adefisan EA**, **Anderson ML**, **Aura S**, **Balogun AA**, **Bertram D**, **Bluhm S**, **Brayshaw DJ**, **Browell J**, **Büeler D**, **Charlton-Perez A**, **Chourio X**, **Christel I**, **Coelho CAS**, **DeFlorio M**, **Delle Monache L**, **di Giuseppe F**, **García-Solórzano AM**, **Gibson PB**, **Goddard L**, **González Romero C**, **Graham RJ**, **Graham RM**, **Grams CM**, **Halford A**, **Huang WTK**, **Jensen K**, **Kilavi M**, **Lawal KA**, **Lee RW**, **MacLeod D**, **Manrique-Suñén A**, **Martins ESPR**, **Maxwell CJ**, **Merryfield WJ**, **Muñoz ÁG**, **Olaniyan E**, **Otieno G**, **Oyedepo JA**, **Palma L**, **Pechlivanidis IG**, **Pons D**, **Ralph FM**, **Reis DS**, **Remenyi TA**, **Risbey JS**, **Robertson DJC**, **Robertson AW**, **Smith S**, **Soret A**, **Sun T**, **Todd MC**, **Tozer CR**, **Vasconcelos FC**, **Vigo I**, **Waliser DE**, **Wetterhall F and Wilson RG** (2022) Advances in the application and utility of subseasonal-to-seasonal predictions. *Bulletin of the American Meteorological Society 103*(6), E1448–E1472. https://doi.org/10.1175/BAMS-D-20-0224.1.

Wilhite DA and Pulwarty RS (2017) Drought and water crises: Lessons drawn, some lessons learned, and the road ahead. *Chapter*, *24*, 513–525.

Wilks DS (2011) *Statistical Methods in the Atmospheric Sciences*, Vol. *100*. Oxford: Academic press.

Cite this article: Pyrina M, Imamovic A, Bueeler D, Spirig C and Domeisen DIV (2026). Joint bias correction and downscaling of subseasonal forecasts via diffusion models. *Environmental Data Science*, 5: e22. doi:10.1017/eds.2026.10047

https://doi.org/10.1017/eds.2026.10047 Published online by Cambridge University Press