Supplementary Material

Joint Bias Correction and Downscaling of Subseasonal Forecasts via Diffusion Models

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

[1]European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany
[2]Center for Climate Systems Modeling (C2SM), Zurich, Switzerland
[3]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[4]Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland
[5]Faculty of Geosciences and Environment, Université de Lausanne, Lausanne, Switzerland

1. Additional Results

The training input of the diffusion models (DMs) consists of ECMWF forecasts interpolated from 36 km to 2 km resolution. This interpolation step can amplify inherent model biases at all lead times, particularly over regions with complex orography such as Switzerland. In addition, the ECMWF forecasts are compared against the observational TabsD dataset that is independent from the data used during model initialization. As a result, discrepancies related to resolution mismatches and the use of independent observational references may contribute more strongly to the overall model error than the lead-time-dependent forecast drift itself. In Figure S1, we present the daily 2t bias for selected Swiss cities for the interpolated ECMWF data. These biases correspond to those shown in Figure 2 of the main manuscript, but are here resolved at daily timescales and across multiple locations to better highlight spatial and temporal patterns. Notably, the bias remains relatively stable for some locations either across all lead times (e.g., Zurich and Geneva in Figure S1) or when comparing shorter (first two weeks) and longer lead times (e.g., Locarno and Sion in Figure S1). The relative stability of the bias at specific locations and/or lead times is consistent with findings reported by Dutra et al. (2021) (see their Figure 3a,b), who evaluated subseasonal ECMWF forecasts against ERA5—the dataset used for ECMWF model initialization (Hersbach et al., 2020; Vitart et al., 2019).

To further assess whether quantile mapping (QM) or DM-based approaches provide better probabilistic forecasts in terms of the spatial structure of the ensemble predictions, we computed the Energy Score (ES; Figure S2). The ES provides a single scalar measure of probabilistic forecast quality, jointly reflecting calibration (statistical consistency with observations) and sharpness (the spread of the predictive distribution). Unlike looking at grid points individually (which would be the CRPS), the Energy Score can evaluate the entire Swiss grid as a vector. It takes into account whether the model correctly predicts spatial relationships, for example, if it correctly forecasts that it will be warm simultaneously in Zurich and Geneva. The results indicate that QM still outperforms the DM approaches in terms of this metric. However, all methods exhibit relatively high ES values. This can likely be attributed to the complex terrain of Switzerland and the very high spatial resolution (2 km) used for evaluation, both of which can introduce substantial spatial mismatches between forecasts and observations, thereby penalizing this probabilistic score.

*maria.pyrina@ecmwf.int

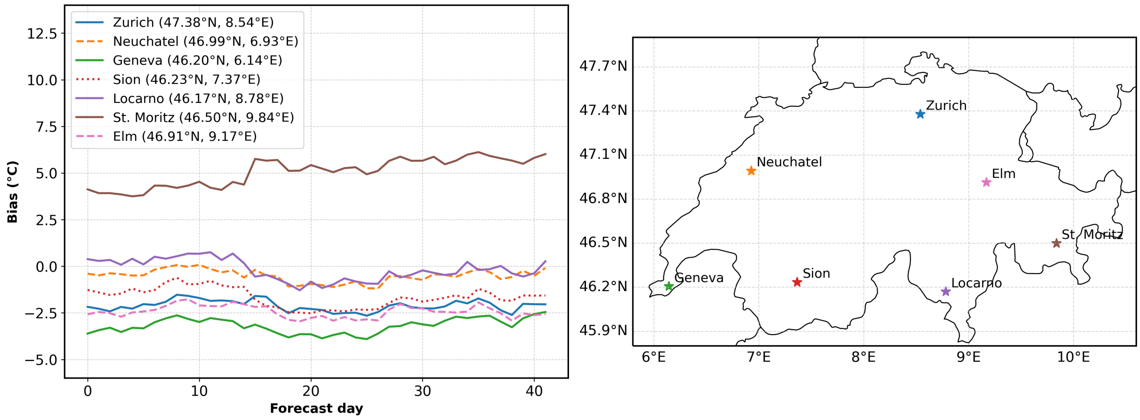


Figure S1: Daily 2t biases for selected Swiss cities, based on the interpolated ECMWF IFS cycle 47r3, are shown for the June–August 2020 period (left panel). When averaged weekly, these correspond to the biases for different forecast lead weeks presented in Figure 2. The map on the right indicates the locations of the Swiss cities considered. Note the jump in bias for forecast day 15, which is related to the change in subseasonal IFS 47r3 native horizontal resolution of approximately 18 km for days 1–15 and 36 km afterwards.

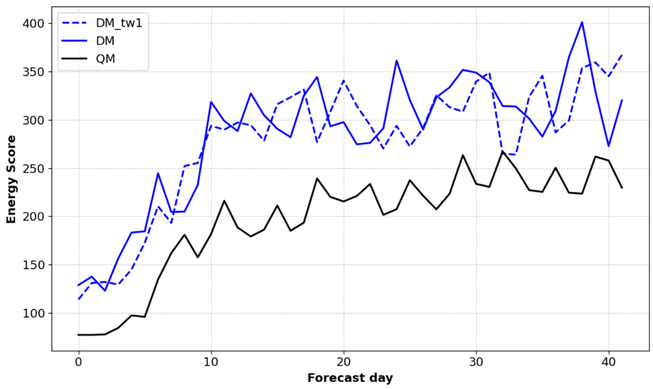
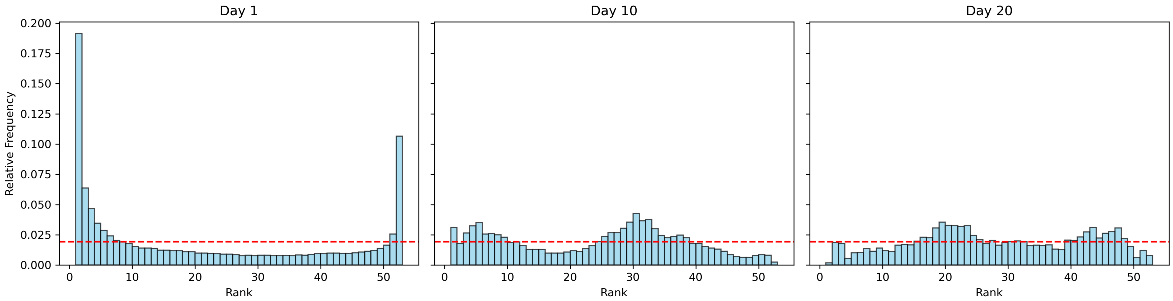
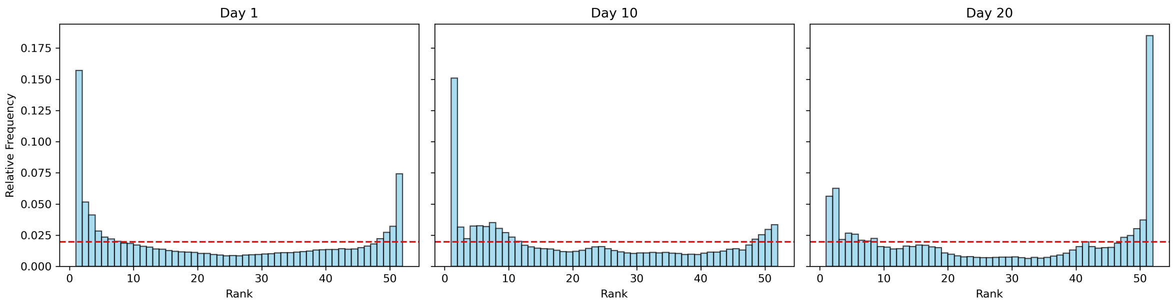


Figure S2: Energy score (ES) of daily 2t ensemble prediction over all model initializations between June-August 2020 for the DM_all approach (blue line), the DM_tw1 approach (blue dashed line), and the QM approach (black line).

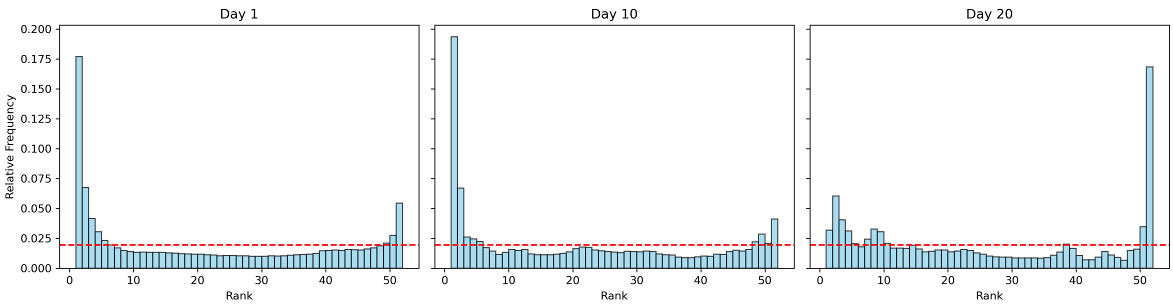
Finally, to better understand differences in ensemble calibration between QM and DM approaches, we analyzed rank histograms (Figure S3). For day 1, both the DM models and QM exhibit underdispersion, indicating that ensemble spread is generally too small. Nevertheless, the DM models perform slightly better in this respect, as evidenced by a lower frequency of observations falling outside the ensemble range. At longer lead times (days 10 and 20), the QM approach shows a multi-modal rank distribution, suggesting the emergence of systematic biases in the ensemble over time. In particular, at day 10, observations tend to cluster both in the middle-to-upper ranks (around rank 30) and in the lower ranks (around rank 5), highlighting deficiencies in representing the full predictive distribution. In contrast, while the DM models remain somewhat underdispersive at day 10, their rank histograms are more uniform across ranks, indicating a more consistent and reliable representation of forecast uncertainty compared to QM.



(a)



(b)



(c)

Figure S3: Rank histograms of 2t ensemble forecasts for day 1, day 10, and day 20, computed over all model initializations, using (a) the QM method, (b) the DM_all approach, and (c) the DM_tw1 approach for the June–August 2020 period. The red dashed line represents a perfectly calibrated ensemble (a uniform distribution).

References

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.

Hersbach, H., Bell, B., Berrisford, P., and et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*.

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 841, ECMWF.