Supplementary Information for: Skillful Data-Driven Subseasonal Soil Moisture Forecasting: Prospects and Limits for Flash Drought Prediction

Noelia Otero[1*], Atahan Özer[1], Miguel-Ángel Fernández-Torres[2], Jackie Ma[1]

[1*]Applied Machine Learning Group, Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany.
[2]Department of Signal Theory and Communications, Universidad Carlos III de Madrid (UC3M), Madrid, Spain.

*Corresponding author(s). E-mail(s): noelia.otero.felipe@hhi.fraunhofer.de;

Supplementary Text 1: Sensitivity to input window length

We assess the sensitivity of SMCast to the length of the input sequence by training models with 60, 90, and 120 days of daily input fields. All configurations use the core SMCast architecture, featuring Rotary Position Embeddings (RoPE), a learnable spatial-temporal fusion gate, and a token embedding dimension of 1024, optimized under the residual formulation with SM100 as the target variable. Table S1 reports the MAE (in normalized units) for each pentad lead time across the three temporal window configurations, alongside the persistence baseline for comparison.

The 90-day window achieves the highest overall skill (with a mean MAE of 0.1689, compared to 0.1712 for the 60-day and 0.1694 for the 120-day track) and is best at the medium-to-extended lead times (pentads 4 and 6–10), which are the most relevant for subseasonal flash drought early warning. The extended 120-day window yields marginal improvements only at the shortest leads (pentads 1–3 and 5), where short-term land-surface persistence dominates predictability, but offers no performance advantage at longer leads while increasing the computational overhead of the attention mechanism (since GPU memory requirements scale linearly with sequence

Table S1 MAE across lead times for different input window lengths. All models use the SMCast architecture (RoPE + fusion gate, embedding dimension of 1024) with the residual formulation and SM100 as the target. Persistence is shown for reference (90-day evaluation set). Bold indicates the best model at each lead time.

Lead	60 days	90 days	120 days	Persistence
1 (d 1–5)	0.1027	0.1035	**0.1005**	0.1081
2 (d 6–10)	0.1285	0.1289	**0.1279**	0.1375
3 (d 11–15)	0.1477	0.1470	**0.1467**	0.1602
4 (d 16–20)	0.1591	**0.1575**	0.1587	0.1760
5 (d 21–25)	0.1697	0.1676	**0.1674**	0.1932
6 (d 26–30)	0.1856	**0.1817**	0.1819	0.2128
7 (d 31–35)	0.1951	**0.1916**	0.1935	0.2316
8 (d 36–40)	0.1979	**0.1935**	0.1944	0.2382
9 (d 41–45)	0.2042	**0.1999**	0.2028	0.2517
10 (d 46–50)	0.2217	**0.2173**	0.2200	0.2787
Mean	0.1712	**0.1689**	0.1694	0.1988

length in the temporal attention pathway). The 60-day window fails to secure the top tier at any individual lead.

Beyond its empirical advantage, the 90-day window is also physically consistent with the memory timescales of root-zone soil moisture, which range from weeks to several months [1, 2]. This roughly seasonal window of context is sufficient to capture the relevant land-surface memory and slow-varying hydroclimatic signals, while longer windows mainly add remote inputs of diminishing predictive value. We therefore adopt the 90-day configuration as default throughout this study, as it offers the best balance between predictive skill at the lead times of interest and computational cost.

Supplementary Text 2: Architectural ablation (RoPE and fusion gate)

Table S2 Architectural ablation of SMCast on the 2021–2022 test period ($T_{\mathrm{in}} = 60$ days, embedding dimension of 1024, residual target). RoPE denotes rotary temporal positional embeddings; FG denotes the learnable fusion gate. Lower MAE/RMSE is better; higher R^2 and skill scores are better.

Variant	MAE	RMSE	R^2	SS_{clim}	SS_{pers}
RoPE + FG (default)	**0.1712**	**0.2610**	**0.5219**	+0.4701	**+0.1409**
RoPE, no FG	0.1736	0.2638	0.5115	+0.4710	+0.1290
no RoPE, FG	0.1715	0.2622	0.5176	**+0.4767**	+0.1394
no RoPE, no FG	0.1728	0.2628	0.5153	+0.4705	+0.1328

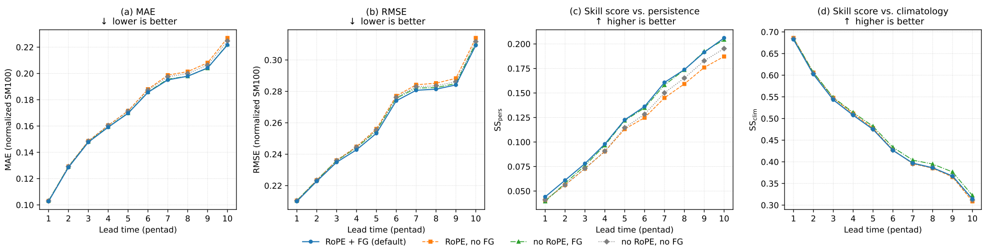

Fig. S1 Architectural ablation study of SMCast across pentad lead times. (a) MAE and (b) RMSE metrics (normalized SM100, lower is better); (c) skill score relative to persistence and (d) skill score relative to climatology (higher is better), for the four configurations combining rotary positional embeddings (RoPE) and the learnable fusion gate (FG). The default configuration (RoPE + FG, highlighted in blue) achieves the strongest and most consistent skill against persistence across all lead times.

Differences between variants are modest in absolute terms, but the integrated RoPE+FG framework delivers the most consistent and well-grounded scores across MAE, RMSE, variance explained (R^2), and skill against persistence (SS_{pers}). Consequently, this framework is selected as the default backbone configuration for both the QH fine-tune and the input-window sequence experiments.

Supplementary Text 3: Deep learning architectures

To ensure a fair comparison between baselines, the models were configured to a parameter count comparable to SMCast ($\sim$65 M parameters); the recurrent E3D-LSTM model is the exception at 81 M due to intrinsic multi-channel recurrent state constraints. Table S3 summarizes the key architectural settings and parameter counts. We note that the U-Net reached its best validation loss substantially faster than the transformer-based models (epoch 11 versus epoch 37 for SMCast), reflecting the strong inductive bias of convolutional encoders for smooth spatial fields; the comparison controls for parameter capacity but not for training compute.

Table S3 Architectural settings and parameter counts of SMCast and the deep learning baselines. All models use 90-day daily inputs (16 channels), output 10 pentad leads, and are trained with the residual formulation and MAE loss.

Model	Key configuration	Params
SMCast (ViT)	embed dim 1024, depth 4, 4 heads, patch 16, decoder depth 3	64.7 M
U-Net	hidden channels 42, 90×16 input channels	65.1 M
CrossFormer	dims [92, 184, 368, 736], depth [2, 2, 8, 2], patch 2	65.5 M
E3D-LSTM	hidden size 16, 1 LSTM layer, τ=30, kernel 1	81.1 M

Supplementary tables and figures

Table S4 reports the per-lead RMSE for the same models, complementing the per-lead MAE results in the main text. The ranking is consistent across both metrics: SMCast achieves the lowest RMSE at every lead time, the U-Net and CrossFormer follow closely, and the recurrent E3D-LSTM remains the weakest data-driven model.

Table S4 Per-lead RMSE (normalized units) on the full 2021–2022 test set, reported as the mean across five random seeds. Conventions as in the per-lead MAE table (main text); the best value at each lead time is in bold.

Lead	SMCast	U-Net	CrossFormer	E3D-LSTM	Persistence	Climatology
1 (d 1–5)	**0.2125**	0.2174	0.2171	0.2587	0.2161	0.4004
2 (d 6–10)	**0.2192**	0.2256	0.2246	0.2682	0.2274	0.3975
3 (d 11–15)	**0.2333**	0.2406	0.2395	0.2847	0.2469	0.3977
4 (d 16–20)	**0.2381**	0.2489	0.2448	0.2953	0.2590	0.3978
5 (d 21–25)	**0.2518**	0.2622	0.2588	0.3130	0.2789	0.3977
6 (d 26–30)	**0.2676**	0.2771	0.2755	0.3339	0.3024	0.3976
7 (d 31–35)	**0.2780**	0.2887	0.2863	0.3497	0.3197	0.3972
8 (d 36–40)	**0.2791**	0.2926	0.2875	0.3592	0.3300	0.3973
9 (d 41–45)	**0.2825**	0.2953	0.2906	0.3701	0.3413	0.3976
10 (d 46–50)	**0.3072**	0.3200	0.3159	0.3979	0.3713	0.3980
Mean	**0.2569 ± 0.0005**	0.2668 ± 0.0068	0.2641 ± 0.0012	0.3231 ± 0.0240	0.2893	0.3979

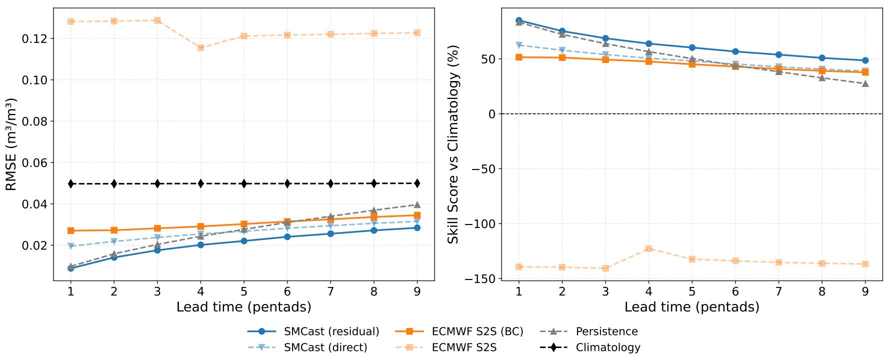

Fig. S2 RMSE (lower is better) across pentad lead times, complementing the MAE comparison in the main text. Different colored lines and shapes correspond to the different models and baselines considered in this study.

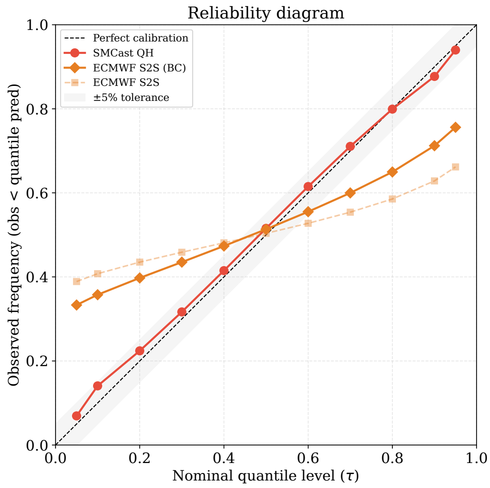

Fig. S3 Reliability diagram evaluating probabilistic calibration. Plot benchmarks SMCast QH (configured for 11 discrete quantiles), ECMWF S2S (raw, 50 members), and ECMWF S2S (BC, 50 bias-corrected members). Points close to the diagonal indicate a well-calibrated probabilistic forecasting framework.

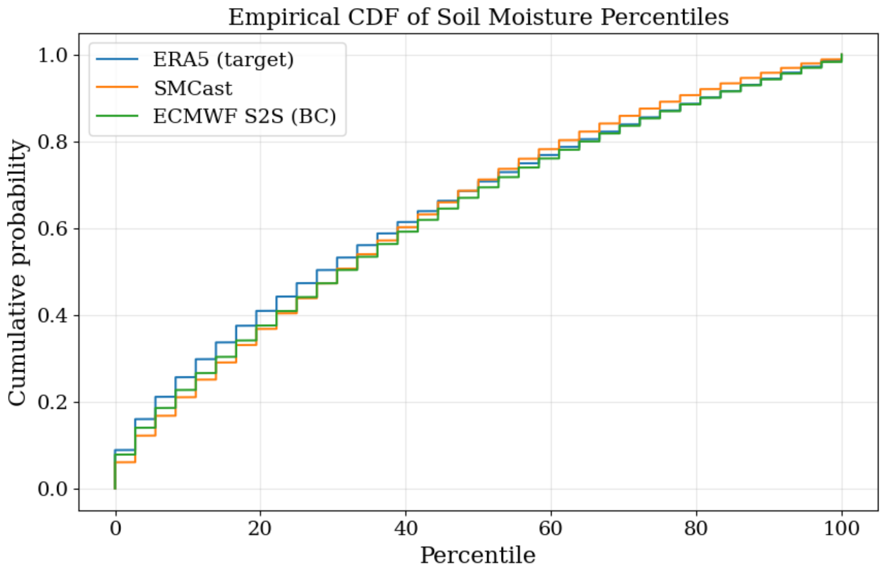

Fig. S4 Empirical cumulative distribution functions (CDF) of SM100 percentiles, restricted to land grid coordinates over the 2021–2022 test period. All three systems show similar distributions, with ERA5 (blue), SMCast (orange), and S2S BC (green) concentrated toward low percentiles, which reflects the severe impacts of the 2022 European drought.

References

[1] Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J.: Investigating soil moisture–climate interactions in a changing climate: A review. Earth-Science Reviews **99**(3), 125–161 (2010) https://doi.org/10.1016/j.earscirev.2010.02.004

[2] Ghannam, K., Nakai, T., Paschalis, A., Oishi, C.A., Kotani, A., Igarashi, Y., Kumagai, T., Katul, G.G.: Persistence and memory timescales in root-zone soil moisture dynamics. Water Resources Research **52**(2), 1427–1445 (2016) https://doi.org/10.1002/2015WR017983