Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Learning skillful forecasting from days to seasons

Siddhant Agarwal

siddhant.agarwal@hereon.de

Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research https://orcid.org/0000-0002-0840-2114

Ali Bekar

Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research

Anthony Frion

Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research

Eduardo Zorita

Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research

David Greenberg

Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research

Research Article

Keywords: subseasonal-to-seasonal prediction, probabilistic forecasting, AI weather models, Earth System modeling

Posted Date: September 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10406658/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Learning skillful forecasting from days to seasons

Siddhant Agarwal[1,3,4*], Ali Bekar[1,4], Anthony Frion[1,4], Eduardo Zorita[2], David Greenberg[1,4]

[1]Model-driven Machine Learning, Helmholtz-Center Hereon, Geesthacht, Germany.
[2]Climate Extremes and Impact, Helmholtz-Center Hereon, Geesthacht, Germany.
[3]Institute of Space Research, German Aerospace Center, Berlin, Germany.
[4]Helmholtz AI, Germany.

*Corresponding author(s). E-mail(s): siddhant.agarwal@hereon.de;

Abstract

Forecasts months ahead support decisions in agriculture, water, energy, and public health, but remain computationally expensive. While artificial intelligence (AI) has transformed medium-range predictions, comparable skill at longer leads remains limited and fragmented across horizons of the same evolving Earth system. Here we show that a single model can learn this evolution from days to seasons with the Functional Graph Transformer (FGT). FGT remains skillful throughout the 14-day weather range and outperforms a specialist subseasonal AI model over weeks 1–6. Across months 1–6, FGT achieves positive skill in 95% of metric–variable–lead combinations and retains at least 80% of operational SEAS5 skill in 76% of cases, establishing broad multivariate seasonal skill. Long-horizon training alone is insufficient: seasonal skill depends on coherent weather and subseasonal evolution, while persisting surface anomalies is worse than omitting them. Free-running 65-year predictions remain stable without explicit budget constraints. These results suggest that physical behavior can be learned from even incomplete representations, but not from inconsistent ones. Learning the dynamics of the underlying Earth system therefore provides a promising route to long-range skill.

Keywords: subseasonal-to-seasonal prediction, probabilistic forecasting, AI weather models, Earth system modeling

1 Introduction

Forecasts several months ahead support decisions in agriculture, water and energy management, public health, and disaster preparedness [1–3]. Although daily weather

loses predictability with lead time due to the chaotic nature of atmospheric dynamics [4], seasonal predictability can emerge from slower variations in the ocean, land, sea ice, and atmosphere. Numerical weather prediction systems such as SEAS5 [5] therefore rely on coupled models and ensembles of simulations to predict distributions of possible outcomes. These systems remain the benchmark for seasonal prediction but are expensive to run and can develop biases for some variables at longer horizons [6, 7]. Artificial intelligence (AI) models have meanwhile transformed medium-range weather forecasting, achieving high skill at a fraction of the computational cost of numerical prediction [8–11]. Extending this advantage to seasonal scales presents an important opportunity, with previous studies broadly falling under two categories.

The first strategy learns and evaluates low-frequency summaries such as climate indices or seasonal aggregates [12–15]. These models are not only limited to a handful of variables such as 2-m temperature and precipitation, but also discard the intervening trajectory that would be necessary to resolve certain phenomena such as individual heatwaves. Coarsening in time to, for example, monthly averages also reduces the number of samples available from reanalysis, motivating training on thousands of years of climate-model simulations. The valuable additional examples come at the cost of an additional source of uncertainty: variability within the climate system is mixed with differences among climate models and their parameterized physics.

The second strategy evolves the system state autoregressively. This is attractive because weather, subseasonal, and seasonal predictions can emerge from the same learned dynamics. Autoregressive AI forecasting without prescribed future surface fields such as sea-surface temperature has already reached subseasonal horizons of up to 6 weeks [16–18]. At seasonal leads of approximately 6 months, however, this route remains incomplete. Existing studies either persist anomalies of slowly varying fields such as sea-surface temperature [19, 20] for December–January–February (DJF) predictions, or demonstrate stable simulations and aspects of climate variability without quantifying skill [21, 22].

This matters because stability is a necessary but insufficient condition for establishing skill. For example, we find that the highlighted 2020 long-range trajectory of Ref. [23] fails to outperform climatology at seasonal leads (Fig. S5). SEAS5, on the other hand, outperforms a probabilistic trend-aware climatology across several components of the Earth system, confirming climatology as a meaningful baseline while simultaneously setting a second, more challenging benchmark for AI models. Furthermore, existing comparisons between probabilistic forecasting systems have often been distorted by unequal ensemble sizes [13, 16], an asymmetry that can strongly affect scores [24] and obscure directions for future improvements by inflating the performance of AI models.

Thus, previous studies leave three questions unresolved. First, can an AI system achieve broad skill at seasonal lead times? Second, can such skill be demonstrated in fair comparisons with climatology and numerical models? Third, although what remains predictable changes across forecast horizons, the underlying Earth system is the same [25]. Can a single system learn these dynamics and let predictability emerge from the bottom up?

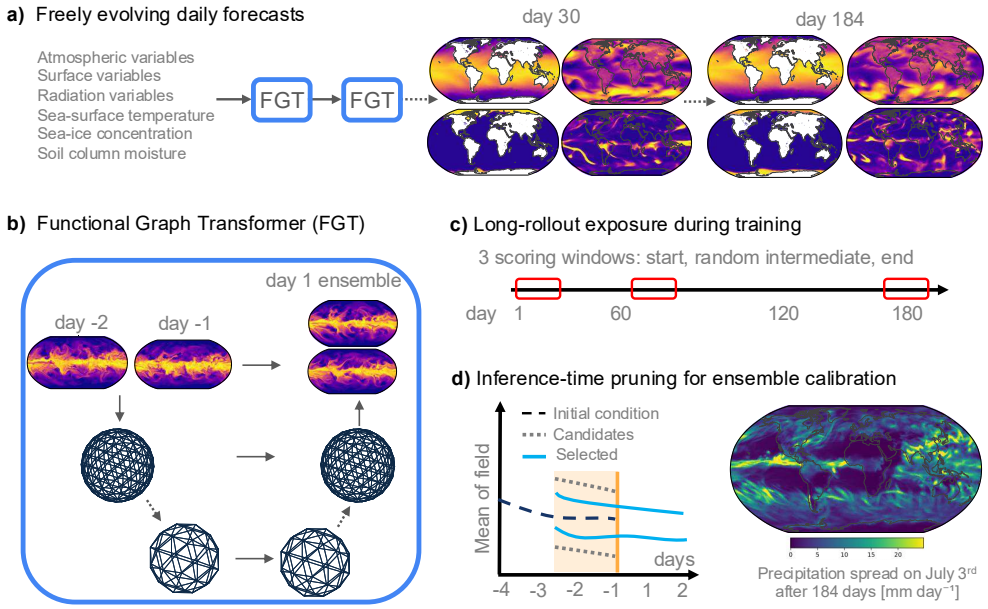


Fig. 1 Overview of the Functional Graph Transformer (FGT) forecasting system. **a)** FGT autoregressively advances daily atmospheric, surface, radiation, sea-surface-temperature, sea-ice, and soil-moisture fields from its own predictions; maps show illustrative forecast fields at days 30 and 184. **b)** At each step, the two preceding daily states are advanced through the attention-based graph U-Net. Each member is conditioned on a distinct 64-dimensional noise vector sampled once and held fixed throughout the rollout. **c)** During long-horizon fine-tuning, FGT is exposed to autoregressive rollouts of up to 180 days. Gradients are retained within three scoring windows located near the start, at a randomly sampled intermediate lead, and near the end of the rollout, training the model on its own states at weather, subseasonal, and seasonal lead times. **d)** At inference, 300 candidate trajectories are generated through a short pre-initialization window and ranked using only states available at initialization. The 51 candidates whose mean-field trajectories most closely match the observed initial-condition trajectory are retained and rolled forward. The map illustrates an example of ensemble spread in daily precipitation after 184 forecast days.

We answer these questions with the Functional Graph Transformer (FGT), a probabilistic autoregressive system that advances daily atmospheric, surface, radiation, sea-surface-temperature, sea-ice, and soil-moisture fields from its own predictions. FGT is trained on self-generated trajectories extending to 180 days, with supervision distributed across weather, subseasonal, and seasonal states using 61 years of ERA5 reanalysis. Temporally coherent probabilistic trajectories are generated by conditioning each rollout on a fixed noise vector, while inference-time ensemble pruning conditions the ensemble on the observed dynamics leading into initialization (Fig. 1).

FGT remains skillful over the first 14 days and outperforms FuXi-S2S over weeks 1–6. Across months 1–6, FGT achieves positive skill over a trend-aware climatology in 95% of metric–variable–lead combinations and retains at least 80% of SEAS5 operational skill in 76% of cases, while running several orders of magnitude faster. Because the evaluation is matched, the remaining gap points directly to subsurface-ocean

dynamics and long-range uncertainty calibration as the most promising directions for improvement.

Ablations show that achieving multihorizon skill within a single system is possible, but not trivial. We find that persisting slowly varying surface anomalies can be more damaging than omitting the fields altogether. Furthermore, seasonal skill depends on maintaining a coherent trajectory through weather and subseasonal states.

Learning the dynamics instead of horizon-specific targets provides a promising route to achieving skill across weather, subseasonal, and seasonal timescales.

2 Results

Unless otherwise noted, forecasts are evaluated against ERA5 over 2020–2026 after model-specific debiasing. The continuous ranked probability skill score (CRPSS) and root-mean-square skill score (RMSS) measure gains over a trend-aware climatology, while spatial and temporal anomaly correlation coefficients (ACC and TCC, respectively) are computed after detrending (Sec. 4.3). Comparisons use matched initialization dates, grids, ensemble sizes, and shared variables.

2.1 FGT remains skillful from days to seasons

FGT was developed for seasonal prediction, but it reaches seasonal lead times by advancing daily states rather than predicting monthly aggregates directly. We therefore ask how the model behaves as it passes through medium-range and subseasonal regimes and beyond the verified forecast horizon (Fig. 2).

Publicly available forecasts provide medium-range references from GenCast [11] and WeatherNext2 (WN2) [26], subseasonal references from FuXi-S2S [16] and IFS [27, 28], and the seasonal benchmark SEAS5. Using these forecast archives preserves each system's intended inference procedure, while matched initialization dates enable direct comparison (Fig. S4; Table S6). At seasonal lead times, recent AI studies do not provide comparable forecast products (Supplementary Note 4.2). To support future comparisons, we release the 51-member FGT forecast archive on Hugging Face [29].

At daily lead times, our model clearly outperforms climatology but is generally less skillful than WN2 and GenCast. This is expected as WN2 and GenCast are higher-resolution specialist weather systems operating at 6- or 12-hour time steps, whereas FGT operates with a one-day time step to support longer rollouts (Table S7). Nevertheless, FGT retains useful short-range variability and skill rather than simply smoothing toward climatology. On the other hand, strong medium-range skill does not by itself imply stable or skillful forecasts over longer autoregressive rollouts. Ref. [30] found that GenCast becomes unstable after approximately 70 days, while in an illustrative six-month rollout we find that WN2 remains competitive for the first few months before its errors increase rapidly (Fig. S6).

At subseasonal lead times, FGT remains more skillful than climatology and achieves higher CRPSS than FuXi-S2S for most variables over weeks 3–6. The advantage is largest for 2-m temperature, sea-surface temperature, and 850-hPa specific humidity, while differences are smaller for large-scale dynamical fields such as mean

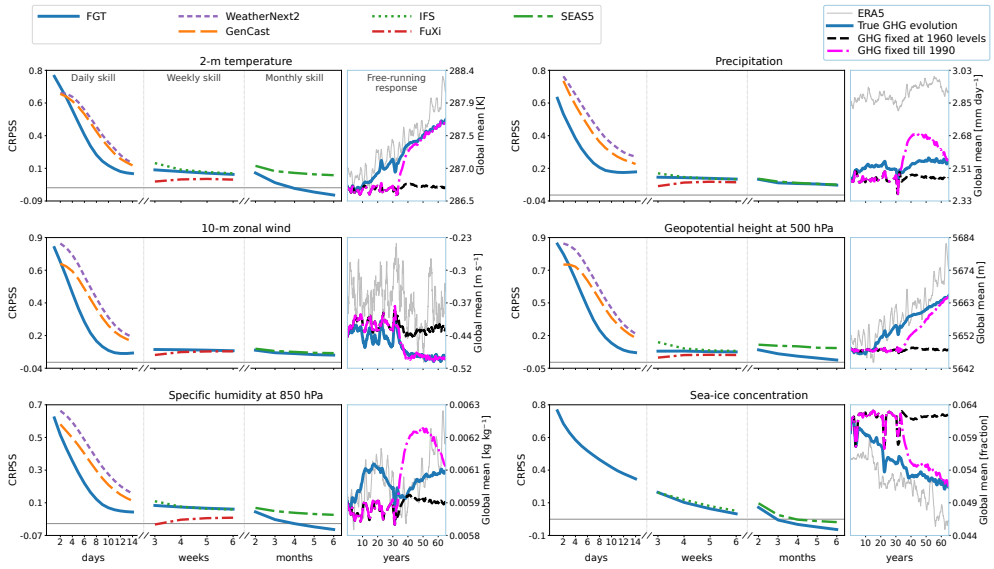


Fig. 2 Multihorizon predictive skill and long-term forced evolution of the Functional Graph Transformer. The same FGT model is evaluated for daily forecasts over days 1–14, weekly means over weeks 3–6, and monthly means over months 2–6 using CRPSS relative to climatology. FGT is compared with WeatherNext2 and GenCast at daily leads, FuXi-S2S and subseasonal IFS at weekly leads, and the 51-member SEAS5 ensemble at monthly leads, using matched initializations. Corresponding RMSS results (Fig. S3) remain positive to longer lead times, showing that deterministic skill over climatology is retained even after the probabilistic gain has been lost. We additionally run FGT freely for 65 years from 1960 under three greenhouse-gas (GHG) trajectories: historical forcing, concentrations fixed at 1960 levels, and concentrations fixed until 1990 before being ramped to present-day levels.

sea-level pressure and geopotential height (Fig. S4). FGT also approaches the skill of the operational 50-member IFS at lead times of approximately 3–4 weeks.

At seasonal lead times, SEAS5 generally remains more skillful, but FGT retains substantial skill while freely evolving its surface states without a dynamical ocean (Sec. 2.3). Our comparison uses 51 members from each system, avoiding the ensemble-size asymmetries of previous AI evaluations [13, 16]. To illustrate the effect of this asymmetry, we compare FGT with a fivefold smaller SEAS5 ensemble (Figs. S12–S16), in which FGT frequently surpasses SEAS5 in deterministic metrics. This highlights a practical advantage of neural forecasting systems: inexpensive large ensembles may outperform numerical models under limited compute or storage budgets. FGT also has lower empirical CRPS than ClimateAR, an AI model that directly predicts monthly states, across all four shared variables and six lead months (Supplementary Table S9).

At multidecadal timescales, beyond the verified forecast horizon, all three 65-year integrations remain bounded without explicit budget constraints or prescribed surface evolution. Under historical GHG forcing, global-mean temperature and 500-hPa geopotential height increase while sea-ice concentration declines, broadly following the ERA5 evolution (Fig. 2). Holding GHG concentrations fixed at their 1960 levels largely suppresses these long-term changes. In the delayed-forcing experiment, the trajectory

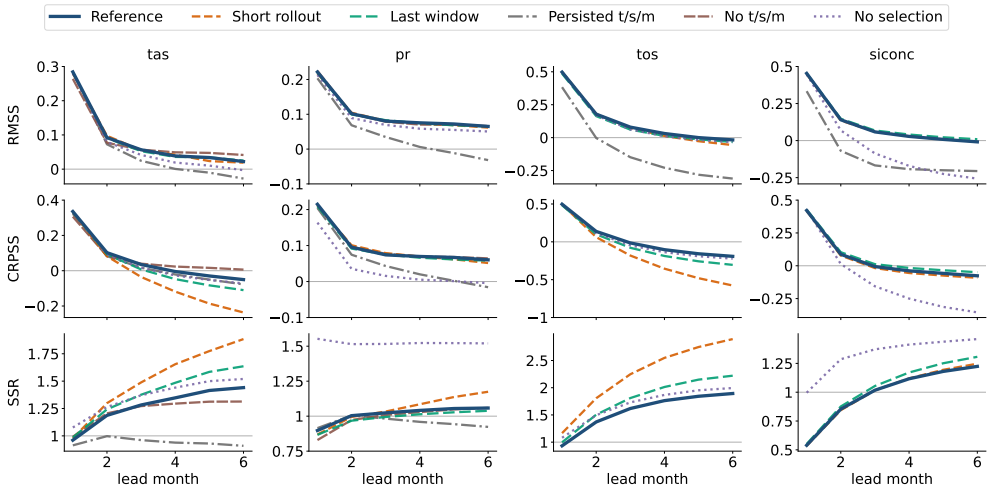


Fig. 3 Forecast skill across lead months for six debiased FGT configurations. The reference model is trained with rollouts of up to 180 steps and retains 51 of 300 stochastic candidates at inference. The short-rollout model uses a maximum training rollout of 15 steps and is given the same training budget as the reference. The last-window model retains 180-step rollouts but applies the loss to one window rather than scoring three windows. Persistence holds anomalies of sea-surface temperature, sea-ice concentration, and soil moisture fixed during inference, whereas the no-surface-fields model excludes them from training and is therefore absent from the corresponding rows. The no-selection model generates 51 trajectories without inference-time candidate selection.

remains close to the fixed-GHG integration until 1990 and then gradually departs as concentrations are increased. Because FGT is trained on a single historical forcing trajectory, the learned response is not independently identifiable [31]. Nevertheless, the stable multidecadal response makes FGT a promising candidate for future training on climate simulations spanning distinct GHG interventions.

These results show that predictability across horizons can emerge from learned Earth-system dynamics, rather than being targeted separately.

2.2 Design choices enabling seasonal skill

We next examine which design choices enable FGT to achieve skill on seasonal scales (Fig. 3). We perform five ablations that isolate rollout length, the temporal placement of supervision, the evolution of slowly varying surface fields, and initialization-conditioned candidate selection. Full scores are reported in Table S5.

Long-horizon exposure improves probabilistic skill and calibration. Relative to the compute-matched short-rollout model, the reference generally maintains higher CRPSS and a spread–skill ratio (SSR) closer to one, particularly for 2-m temperature and sea-surface temperature, and more modestly for soil moisture. Differences in RMSS are smaller. Precipitation is comparatively insensitive to shorter rollout length in RMSS and CRPSS, although it becomes increasingly overdispersed at longer leads.

Distributed supervision is more effective than scoring only the last window of long trajectories. Concentrating the loss at the end hurts CRPSS and increases overdispersion for several surface fields, most clearly 2-m temperature, sea-surface temperature, and soil moisture. This suggests that seasonal skill is not an isolated property to target directly, but can emerge from coherent evolution of the Earth system through weather and subseasonal regimes. As training high-dimensional autoregressive emulators over long horizons can be prohibitively memory-intensive, our distributed scheme provides a practical alternative: different examples in a batch expose the same model to different horizons, without requiring every example to retain gradients through the full rollout.

Freely evolving fields are preferable to persisted anomalies. Under persistence, sea-surface temperature, sea-ice concentration, and soil moisture have zero ensemble spread by design and substantially lower RMSS; feeding the persisted anomalies back into the model also harms other variables, most notably precipitation. In contrast, training from scratch without these variables leaves precipitation skill nearly unchanged and makes the model competitive, or even more skillful, for 2-m temperature and mean sea-level pressure at longer leads. The skill degradation under persistence therefore points to train–inference distribution shift rather than missing physical information: as the remaining variables evolve while the surface anomalies remain fixed, the joint state becomes increasingly incoherent. Excluding these variables during training instead requires the network to learn from the remaining fields.

Candidate selection conditions the ensemble on the dynamics of the initial state and improves calibration. It reduces overdispersion and increases CRPSS for precipitation and several surface, moisture, and radiation variables, while changes in deterministic skill are smaller. Notably, sea-ice concentration improves despite not being included in the selection score (Sec. 1.5), likely through implicit coupling to other fields.

Together, these experiments show that seasonal skill is unlikely to emerge simply by extending inference from a short-range weather model, even when its forecasts remain stable. It helps to train on self-generated states at long horizons, supervise across multiple forecast regimes, and condition probabilistic inference on the initialization dynamics.

2.3 Seasonal forecast skill of FGT

We now focus on the seasonal range for which FGT was designed, where other AI models have so far demonstrated skill mainly for a few variables, seasons, indices, or monthly aggregates. We evaluate global seasonal skill across 56 surface and pressure-level fields (Fig. S10), with representative results summarized in Fig. 4. Across four metrics, FGT achieves significantly positive skill in 1,282 of 1,344 metric–variable–lead combinations (95%). Skill declines with lead time, with CRPSS generally decaying faster than RMSS, TCC, and ACC.

Precipitation, mean sea-level pressure, surface winds, soil moisture, and radiative fluxes remain skillful through month 6, whereas CRPSS for 2-m temperature, sea-surface temperature, and sea-ice concentration becomes negative at later leads. Their correlations and RMSS nevertheless generally remain positive, indicating that FGT continues to capture deterministic skill and anomaly structure after its probabilistic

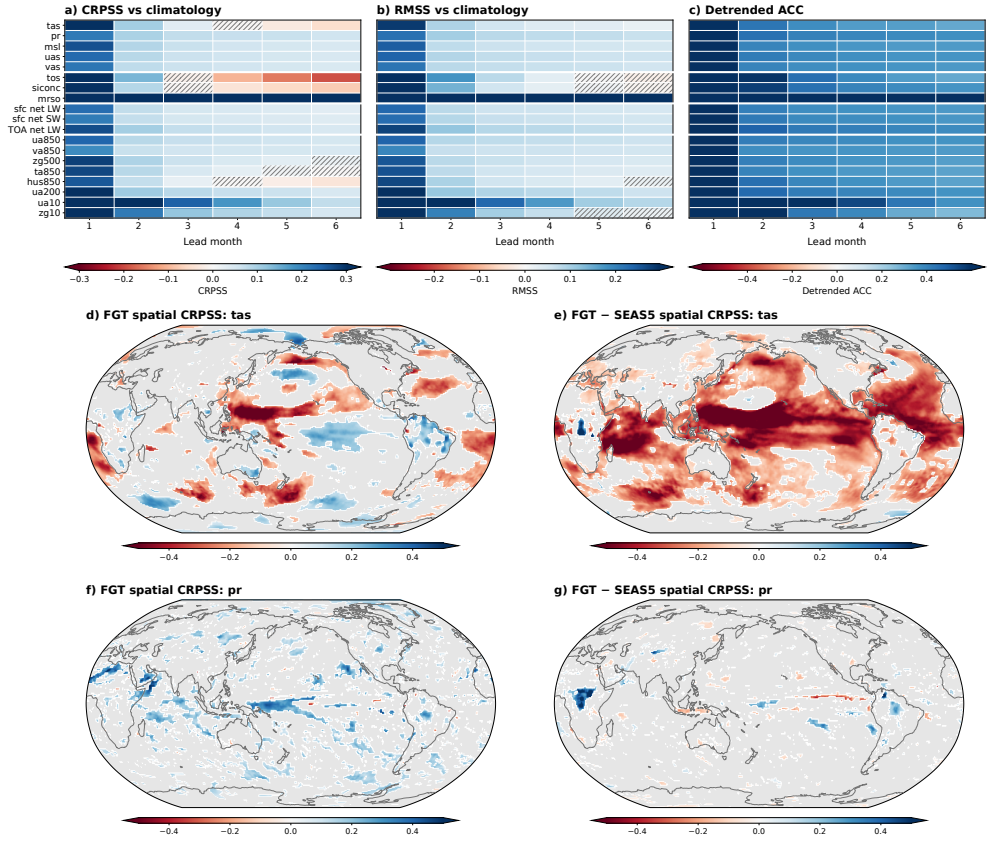


Fig. 4 Global and spatial seasonal forecast skill of the 51-member FGT ensemble over the 2020–2026 test period. **a–c)** Globally averaged skill across representative variables and lead months: **a)** CRPSS and **b)** RMSS relative to the trend-aware climatological ensemble, and **c)** detrended ACC with ERA5. Hatched cells are not significantly different from zero at the 5% level based on bootstrap resampling over initialization dates. **d–g)** Spatial CRPSS at lead month 4 for **d,e)** 2-m temperature and **f,g)** precipitation. Panels **d,f)** show FGT relative to climatology, while panels **e,g)** show the difference between FGT and SEAS5 over matched initialization dates. Positive values (blue) favor FGT relative to the corresponding reference; negative values (red) favor climatology or SEAS5. Grey regions are not significantly different from zero at the 5% level.

advantage over climatology is lost. At pressure levels, winds retain broad skill, while skill is higher for geopotential height in the lower troposphere and for atmospheric temperature around 200–100 hPa.

Among the 52 fields shared with SEAS5, 92% of metric–variable–lead combinations retain more than half of the corresponding SEAS5 skill, 76% retain more than 80%, and 56% retain more than 90% (Fig. S11). Retention is strongest for mean sea-level pressure and horizontal winds, while the largest losses occur for late-lead near-surface and sea-surface temperature and parts of the atmospheric profiles. Long-term changes in near-surface temperature may allow the model to reduce its loss by predicting the

evolving background state while underweighting the interannual anomalies used to assess seasonal skill.

Spatially, precipitation remains more skillful than climatology over fragmented regions, particularly in the tropics, while near-surface zonal wind retains broad positive skill (Fig. 4d–g; Figs. S8 and S9). Temperature CRPSS is widespread at month 2 but declines thereafter, with SEAS5 increasingly favored over tropical and oceanic regions.

An important exception is the tropical Pacific, including parts of the Niño 3.4 region, where skill persists to longer leads, suggesting that FGT captures part of a predictable mode associated with the El Niño–Southern Oscillation (ENSO) despite having no explicit long-term memory.

Overall, FGT is closest to SEAS5 for precipitation and circulation and remains weakest for temperature, particularly sea-surface temperature. FGT is therefore not a more skillful replacement for SEAS5, but the maps show where the computationally cheaper model is already competitive or even surpasses SEAS5. For context, a 51-member ensemble of six-month daily forecasts at 1° resolution can be generated in approximately 19 minutes on a single NVIDIA A100 GPU. Extrapolating a coupled IFS–NEMO benchmark implies 9×10^5 CPU-core-hours for an analogous ensemble [32]. Even assuming perfect parallel scaling, matching FGT's wall-clock time would require approximately 2.8 million CPU cores operating simultaneously.

2.4 Forecasts of seasonal indices

We further examine leading modes of seasonal variability in Fig. 5. FGT retains high Niño 3.4 anomaly correlation through lead month 4 and positive correlation at month 6, but increasingly damps the anomaly amplitude, most clearly during the 2023–2024 El Niño. It therefore preserves much of the timing and sign of ENSO evolution while underestimating its magnitude; SEAS5 better retains both correlation and amplitude.

The Dipole Mode Index (DMI) measures the zonal sea-surface-temperature anomaly contrast between the western and eastern tropical Indian Ocean. FGT's DMI correlation decreases sharply after the first lead month and continues to deteriorate toward month 6. Enriching the forecast state with ocean heat content could improve DMI prediction [33].

The NAO constitutes the clearest limitation of both forecast systems. FGT reaches an ACC of approximately 0.45 at lead month 2, but its correlation becomes nearly zero or negative thereafter. SEAS5 exhibits a similar rapid loss of skill.

As an additional diagnostic, we evaluate FGT forecasts of Niño 3.4, DMI, and DJF NAO over the historical period 1983–1989 using the same fixed climatological calibration as in the main analysis. We train a separate FGT model for which these years are held out. Its similar qualitative behavior indicates that the results observed during 2020–2026 are not unique to the particular sequence of events in that period. Additional analysis finds similar error-based metrics for the two periods in Fig. S17.

For Niño 3.4 and DMI, FGT outperforms persistence of the previous month's anomaly, confirming persisted anomalies as an unnecessarily conservative approximation for autoregressive models.

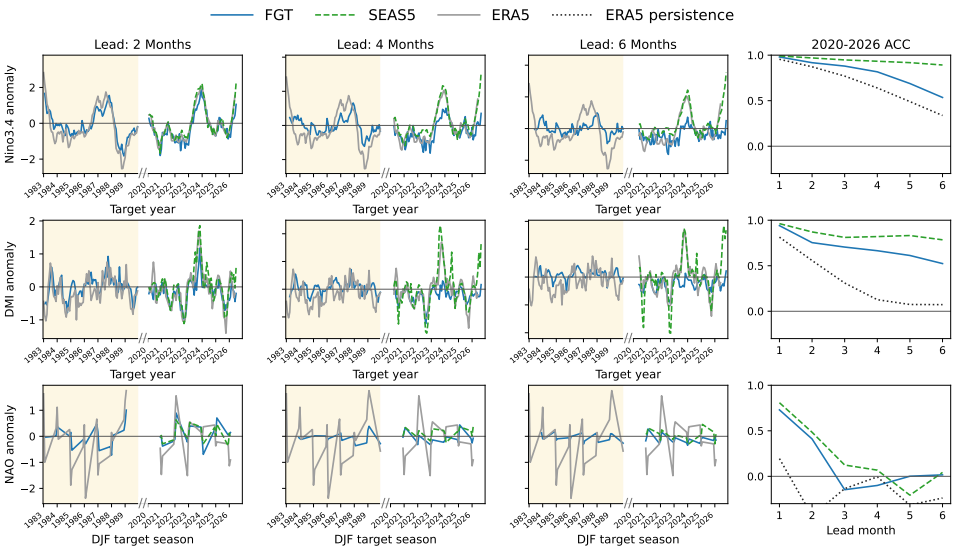


Fig. 5 Seasonal index forecasts over 2020–2026 for Niño 3.4, Indian Ocean Dipole Mode Index (DMI), and North Atlantic Oscillation (NAO). Lead-month panels compare ensemble-mean FGT (solid blue) and SEAS5 (dashed green) forecasts with ERA5 (solid grey); the right column summarizes ACC by lead month over common valid initialization dates and further includes persistence of the previous month's ERA5 anomaly (ERA5Persist; dotted black). The standardized monthly NAO index is shown separately for December, January, and February (DJF). At each lead, ACC is computed by pooling the three monthly values across all complete DJF seasons. As a further diagnostic, we include FGT index forecasts for 1983–1989. We note that these are evaluated with a separately trained FGT for which the years 1983–1989 are unseen.

2.5 Probabilistic forecasts of selected high-impact events

Seasonal forecasts are often used to assess the risk of unusual regional conditions. A comprehensive evaluation of extremes with a much larger event set is beyond the scope of this study. Here we instead use four high-impact events as examples: the August 2020 Indian heavy-rainfall event, the February 2021 North American cold wave, the March–April–May (MAM) 2022 East Africa drought, and the July 2022 European heat event. For each case, Fig. 6 compares the forecast distribution of the regional anomaly with the historical ERA5 distribution and shows the corresponding ensemble-mean anomaly maps. Distributions were estimated using Gaussian kernel density estimation with bandwidth selected according to Silverman's rule [34].

For the August 2020 Indian heavy-rainfall event, the May-initialized FGT forecast places more probability on wetter-than-normal conditions and has a CRPSS of +0.51, compared with +0.08 for SEAS5. The FGT ensemble mean also has some agreement with the observed rainfall pattern, with an ACC of 0.39, while the SEAS5 ensemble mean has little spatial agreement with the observed pattern (ACC = 0.05).

Neither system successfully predicts the February 2021 United States cold wave from the November initialization. Both forecasts therefore have negative CRPSS,

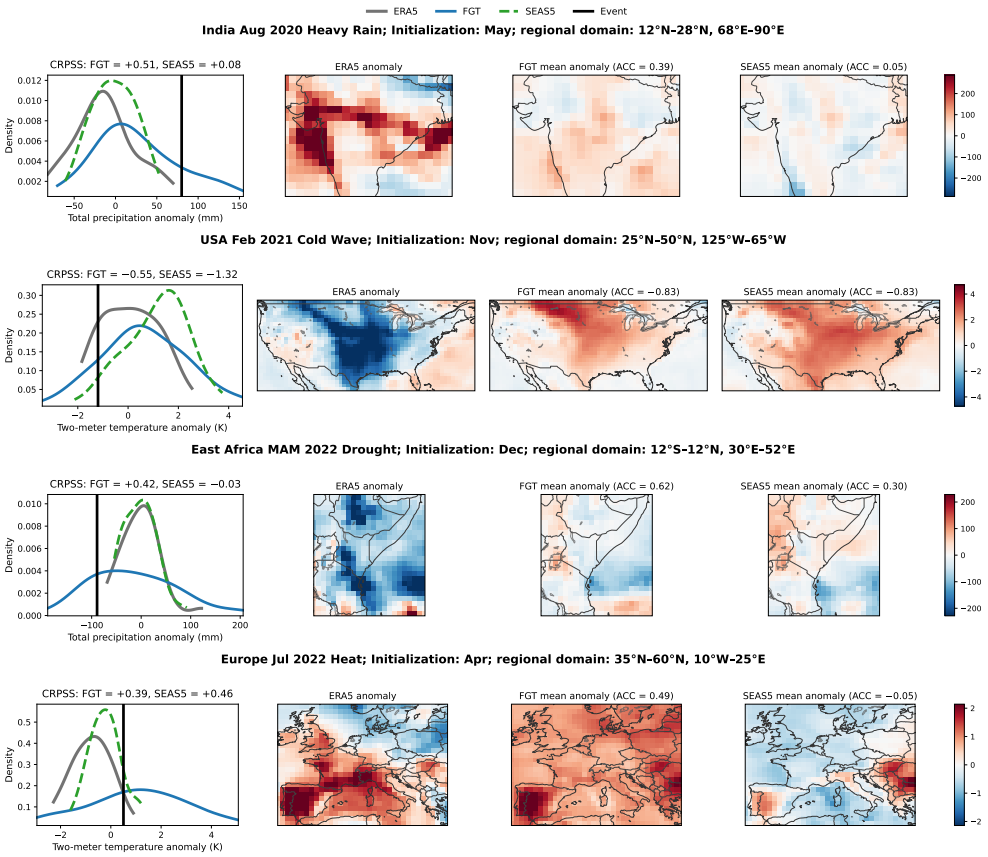


Fig. 6 Regional anomaly distributions and ensemble-mean anomaly maps for four selected high-impact events. Each row shows one event at one initialization. In the left panels, the curves show the distributions of regional anomalies from FGT (solid blue) and ECMWF SEAS5 (dashed green), together with the historical ERA5 anomaly distribution (solid grey). The black vertical line marks the realized ERA5 event anomaly, and the annotations give the CRPSS of each forecast relative to the historical ERA5 distribution. The remaining panels show the realized ERA5 anomaly and the corresponding FGT and SEAS5 ensemble-mean anomalies. Forecast-map titles report the detrended spatial ACC with ERA5. Precipitation anomalies are shown for the India and East Africa events, and 2-m temperature anomalies for the United States and Europe events.

although FGT is less negative than SEAS5 (−0.55 compared with −1.32). For an initialization on 1 February, however, FGT captures the cold anomaly that began forming on 6 February, whereas SEAS5 still shows little spatial agreement with ERA5 (Fig. S2).

For the MAM 2022 East Africa drought, the December FGT forecast assigns more probability to negative rainfall anomalies than either the historical distribution or SEAS5. Its CRPSS is +0.42, whereas the SEAS5 score is close to zero. The FGT ensemble mean also captures a substantial part of the observed dry pattern, with an ACC of 0.62. SEAS5 shows a weaker ACC of 0.30.

For the July 2022 European heat event, both systems improve on the historical reference for the regional anomaly. SEAS5 has a slightly higher CRPSS than FGT (+0.46 compared with +0.39). The spatial correlations of the ensemble means give a different result. FGT captures the broad warm conditions over Europe, although the forecast warmth is more widespread than observed, and reaches an ACC of 0.49. The SEAS5 ensemble mean has little correspondence with the observed spatial pattern and has an ACC of −0.05.

These examples suggest that FGT ensembles contain useful information about the sign of a regional anomaly and its broad spatial pattern. These illustrative cases should be considered alongside the broader test-set results (CRPSS and ACC here), but do not point to a distinct failure mode in forecasting extreme events, consistent with recent evidence [35].

3 Discussion

We have introduced the Functional Graph Transformer (FGT), a probabilistic forecasting system that advances daily-mean atmospheric and surface fields from its own predictions. To our knowledge, FGT is the first AI model to achieve broad, quantifiable seasonal skill while freely evolving sea-surface temperature, sea ice, and soil moisture. Across lead months 1–6, FGT achieves significantly positive skill scores or anomaly correlations in 95% of evaluated metric–variable–lead combinations and, under a matched evaluation, retains at least 80% of operational SEAS5 skill in 76% of cases. A 51-member, six-month FGT ensemble can be generated in 19 minutes on a single GPU. Even under perfect parallel scaling, the coupled IFS–NEMO would need approximately 2.8 million CPU cores to match this wall-clock time. FGT therefore substantially lowers the computational barrier to probabilistic seasonal forecasting.

The significance of this result lies not only in the seasonal skill FGT achieves, but also in how it achieves that skill: by advancing the same Earth-system state through weather, subseasonal, and seasonal regimes rather than targeting seasonal quantities directly. FGT retains useful skill at medium-range lead times, outperforms the specialist FuXi-S2S model over weeks 1–6, and remains stable in 65-year free-running integrations.

Such behavior is not obtained simply by extending inference: long autoregressive rollouts of weather models can develop rapidly increasing errors at longer leads (Fig. S6), or remain stable without skill (Fig. S5). Distributed supervision at weather and subseasonal leads improved seasonal performance, showing that long-range skill depends on coherent evolution of weather trajectories. Persisting slowly varying surface anomalies is more damaging than omitting these variables entirely, indicating that the model can accommodate missing information during training more readily than dynamically inconsistent states at inference.

The 35-million-parameter FGT achieves substantially better probabilistic performance (lower CRPS) than a 480-million-parameter model that directly predicts monthly averages [15]. FGT is trained only on 61 years of ERA5, while ClimateAR is pretrained on nine CESM2 simulations spanning 1,485 simulated years and then fine-tuned on 42 years of ERA5 and ORAS5. Beyond differences in architecture and

training, learning at daily time steps may also make the problem easier: each prediction step is shorter, and the same ERA5 record provides many more training examples. This result challenges scaling-first and horizon-specific views of emulating complex dynamical systems (Table S9).

Because the evaluation against SEAS5 is matched, the remaining performance gap identifies a clear direction for further development. First, FGT learns atmosphere–surface relationships through a shared representation, as do the surface–ocean AIFS and GraphDOP at medium-range lead times [36, 37], whereas Ola and DLESyM represent the atmosphere and ocean with separate learned components [21, 22]. Our results do not establish that either coupling strategy is inherently superior. Instead, the deterioration of sea-surface-temperature skill suggests that surface-only ocean evolution is likely insufficient and that explicit coupling may better accommodate disparate timescales and subsurface-ocean dynamics. A natural extension would therefore be to couple FGT's atmosphere–surface operator to a dedicated ocean component representing subsurface temperature, upper-ocean heat content, and sea-surface height [38]. Such coupling may require correction of inconsistencies in surface energy and momentum fluxes, as in the early development of coupled climate models [39]. A recurrent latent-state memory could additionally support larger time steps with a limited set of variables [33, 40].

Second, FGT ensembles are sometimes overdispersed, pointing to probabilistic inference as another target for improvement. The present inference-time compute strategy provides only a limited mechanism for improving calibration. Future work could explore adapting a small subset of model parameters to initial conditions.

Third, FGT shares a broader limitation of AI models trained on ERA5: the small number of independent climate events available for training and testing and the single GHG scenario it provides. The 2020–2026 test period nevertheless includes substantial variability, including the prolonged La Niña and the strong 2023–2024 El Niño. Furthermore, hindcasts for held-out data in 1983–1989 show similar error behavior (Fig. S17). Training future versions on Earth system model simulations could expose the model to more climate states and forcings and enable evaluation on some additional years [13, 14, 41]. This extension is non-trivial, however, because it requires reconciling model-dependent dynamics, as well as heterogeneous temporal aggregation, particularly for ocean variables often available in public archives only as monthly means.

Overall, FGT shows that AI models can do more than remain stable: they can learn useful dynamics from days to seasons, opening a new frontier in Earth-system prediction.

4 Methods

4.1 Dataset

We use ERA5 daily-mean fields [42] on a regular 1° latitude–longitude grid with 180×360 grid cells. Data are obtained from WeatherBench2 [28]. For dates from 2023 onward that are not covered by WeatherBench2, we construct the same fields from the Copernicus Climate Data Store [43, 44].

All fields are remapped to 1° horizontal resolution using local first-order conservative overlap weights. For the CDS-derived extension, radiative flux variables are handled separately from the other variables: non-radiative surface variables and pressure-level variables use UTC daily means, while radiative fluxes use a UTC+01 daily window to match the WeatherBench2 valid-time convention. Radiation accumulations are converted to mean fluxes in W m^{-2}. Training years are 1959–2019 and test years are 2020–2026. Fields are standardized channel-wise using training-set means and standard deviations from ERA5. Table S1 shows the variables used in this study. Following recent AI models, which treat accumulated precipitation as a diagnostic rather than a prognostic variable, we predict precipitation but do not feed it back to the network [11, 26].

The model is further conditioned on six scalar forcing values: mole fraction of carbon dioxide in air, mole fraction of methane in air, mole fraction of nitrous oxide in air, stratospheric aerosol optical depth at 550 nm, and two deterministic annual-calendar encodings given by the sine and cosine of the day of year. The greenhouse-gas mole fractions were taken from the CMIP6 input4MIPs historical greenhouse-gas concentration dataset documented by Ref. [45]; for dates beyond the forcing record, these greenhouse-gas series were extended using linear extrapolation of recent trends. Stratospheric aerosol optical depth at 550 nm was obtained from the NASA GISS dataset. The four-layer zonal-mean field was summed vertically and then averaged across latitude using cosine-latitude area weights to produce a single scalar forcing channel [46]. These forcing channels are included to distinguish the evolving forcing regimes within the 1959–2019 training period and to support future training on CMIP scenario simulations.

4.2 The Functional Graph Transformer forecasting system

4.2.1 Architecture

FGT is a U-Net-type [47] graph transformer architecture. Let $X_{i,t}$ denote the prediction for ensemble member i at time t. At each autoregressive step, the model $\mathcal{F}_\theta$ receives the two most recent states concatenated in channel space, $[X_{i,t-1}, X_{i,t}]$, and predicts the next state:

$$X_{i,t+1} = \mathcal{F}_\theta \left(X_{i,t-1}, X_{i,t}, F_{t+1}, S, z_i \right). \tag{1}$$

Here F_{t+1} denotes forcing channels such as greenhouse gas concentrations, S denotes static spatial inputs such as orography, SIREN-based sinusoidal embeddings of spherical coordinates [48], and a channel-wise binary validity mask. The mask is one everywhere for globally defined variables, one over ocean and zero over land for ocean variables, and one over land and zero over ocean for land variables. $z_i \in \mathbb{R}^{64}$ is the stochastic noise vector for ensemble member i. At the first time step, $X^{\mathrm{obs}}_{i,t-1}$ and $X^{\mathrm{obs}}_{i,t}$ represent the initial conditions from ERA5.

Input fields are defined on a regular latitude–longitude grid and are projected by the encoder to four progressively coarser icosahedral meshes containing 40,962, 10,242, 2,562, and 642 nodes, respectively. The decoder mirrors this hierarchy back to the

latitude–longitude grid. FGT does not advance a compressed latent representation, as learned latent representations alone may sacrifice spatial inductive biases [49]. Therefore, FGT uses the graph U-Net itself as the autoregressive backbone. Skip connections concatenate encoder and decoder features at every mesh level, providing pathways for both large-scale and fine-scale information during rollout.

We do not predict a residual update to be added to the previous state, as in GraphCast and GenCast [10, 11]. Explicit residual updates require the model to learn not only how new structures emerge, but also how features from the previous time step should be removed. This can place a substantial burden on the network during training. AIFS [50] mitigates this issue by coarsening the previous state before applying a residual update. In addition, residual predictions can lead to the accumulation of high-frequency artifacts during long autoregressive rollouts [51, 52]. Thus, FGT predicts the next state directly and lets skip connections across the network hierarchy learn updates at multiple spatial resolutions.

We use self-attention layers at each icosahedral level and pooling attention between adjacent mesh resolutions, including transfers between the regular latitude–longitude grid and the icosahedral mesh. We use the icosahedral graph attention introduced by the Appa model [53]. Attention is sparse: each node attends only to nodes connected by a local graph-hop neighborhood. Pooling and most self-attention layers use 3-hop neighborhoods, except for self-attention on the coarsest mesh, where the neighborhood is widened to 8 hops to increase the receptive field.

Ensemble generation follows the functional noise perturbation approach of Ref. [26]. Each ensemble member is conditioned on a sampled Gaussian vector z_i with zero mean and unit variance, which is embedded with a multilayer perceptron (MLP). The resulting noise embedding modulates attention blocks through adaptive layer normalization (AdaLN) [54, 55]. We hold the perturbation fixed for each member over the rollout and found that this results in more temporally coherent trajectories than resampling the perturbation at each time step. Furthermore, for stability, the noise embedding is multiplied by a learned scalar, and the AdaLN scale, shift, and residual gates are bounded with tanh nonlinearities. To condition the network on scalar forcings, we embed F_{t+1} with an MLP. In the encoder, this embedding is added to the normalized self-attention features, conditioning the interpretation of the input states. In the decoder, forcing embeddings are concatenated with the node features to inform the prediction of the next state [56].

4.2.2 Model training

The model is trained autoregressively with an almost fair CRPS (afCRPS) objective [50] computed on the normalized fields. The afCRPS is a combination of the continuous ranked probability score (CRPS), which is a strictly proper scoring rule [57], and its fair variant [58], the fair CRPS [59]. For an ensemble forecast $\{X_{i,t+1}\}_{i=1}^{E}$ and target state Y_{t+1},

$$\text{afCRPS} = \frac{1}{E}\sum_{i=1}^{E}|X_{i,t+1} - Y_{t+1}| - \frac{E-1+\alpha}{E^2(E-1)}\sum_{i<j}|X_{i,t+1} - X_{j,t+1}|\,, \tag{2}$$

where $\alpha \in (0, 1]$ controls the finite-ensemble adjustment. We use $\alpha = 0.8$ to slightly reduce the reward for ensemble dispersion and encourage sharper ensemble forecasts. This pointwise loss is averaged over grid cells and channels using the validity mask, cosine-latitude weights, and channel weights λ_c. We set $\lambda_c = 1$ for 2-m temperature and $\lambda_c = 0.1$ for all other channels. Training proceeds in four stages:

1. We first train with a rollout curriculum that increases the horizon up to 8 steps using 4 ensemble members.
2. We then fine-tune on a 15-step rollout with 4 ensemble members.
3. Next, we use a mixed rollout–ensemble regime that samples different rollout/member combinations, reaching rollouts up to 30 steps and ensembles up to 6 members (due to GPU memory constraints).
4. In the final stage, we use a pushforward-style windowed backpropagation scheme for long-horizon fine-tuning [60]. The model is rolled out autoregressively over increasing rollout horizons, exposing it to its own long-range forecast states. We increase the horizon from 30 to 180 in increments of 5 approximately every 25 training steps. To keep memory manageable, we avoid backpropagating through the full trajectory. Gradients are retained only within three selected scoring windows of equal size: an initial window, a randomly sampled intermediate window, and a tail window.

4.2.3 Model inference

At inference, candidate ensemble members are generated by sampling multiple noise vectors and running a short pre-initialization spin-up. Let $t = t_0$ denote the forecast initialization time. Starting from the observed states $[X^{\text{obs}}_{i,t-3}, X^{\text{obs}}_{i,t-2}]$, each candidate first predicts $X_{i,t-1}$ and then uses $[X^{\text{obs}}_{i,t-2}, X_{i,t-1}]$ to predict $X_{i,t}$. Candidate trajectories are ranked by their agreement with the observed states $[X^{\text{obs}}_{i,t-1}, X^{\text{obs}}_{i,t}]$, and the closest members are retained for rollout beyond t_0. Thus, member selection uses only observations available at forecast initialization, and no information from target states after t_0 is used. Related strategies have been used in analogue-based stochastic forecasting [61, 62] and in observation-conditioned sub-selection of lagged ensemble members [63]. As noted by Ref. [63], such selection can be interpreted as a simple form of data assimilation, because observations available at initialization are used to condition the forecast ensemble [64].

4.2.4 Rollout sanitization

During both training and inference, each predicted state is sanitized before it is used as input for predicting the next time step. Values outside the spatial domain of a variable are set to zero in normalized model space using the channel-wise validity mask. We also apply limited physical bounds: sea-ice concentration is clipped to $[-10^{-6}, 1.01]$, while precipitation and total soil moisture are lower-bounded at $-10^{-3}\,\text{mm}\,\text{day}^{-1}$ and $-1\,\text{kg}\,\text{m}^{-2}$, respectively; we found that these small tolerances improved training stability compared with strict physical bounds. During training, the loss is evaluated on the unclipped prediction, so clipping affects the state propagated through the rollout. At inference, the sanitized prediction is retained and propagated to the next step.

We report further details on the architecture hyperparameters and training curriculum in Sec. 1.

4.3 Evaluation

Forecasts are evaluated in anomaly space. Let $X^{\mathrm{raw}}_{i,t}$ denote ensemble member i at target time t, and let Y_t be the verifying ERA5 state. We debias each forecast using the model climatology M_t and the ERA5 climatology C_t:

$$X^{\mathrm{db}}_{i,t} = C_t + \left(X^{\mathrm{raw}}_{i,t} - M_t\right). \tag{3}$$

Here M_t and C_t are fields computed for the relevant initialization time, lead time, channel, and grid point. M_t is determined from each model's hindcast. For consistency with the SEAS5 bias estimate $M_{t,\,SEAS5}$, we compute FGT's climatology $M_{t,\,FGT}$ using 25 members over a prediction period of 1993–2016. The anomaly fields used for RMSE, CRPS, and spread are

$$x_{i,t} = X^{\mathrm{raw}}_{i,t} - M_t, \qquad y_t = Y_t - C_t, \qquad \bar{x}_t = \frac{1}{E}\sum_{i=1}^{E} x_{i,t}. \tag{4}$$

All global averages use channel-validity masks (see Sec. 4.2.1) and cosine-latitude weighting. For sea-ice concentration, scores are computed only over a validity mask containing grid cells where it exceeds 0.15 in ERA5 at least once in that calendar month over the training period, dilated by two grid cells; the 0.15 threshold follows the standard sea-ice-edge definition used in seasonal sea-ice forecast verification [65, 66]. For any gridded quantity A, $\langle A\rangle_w$ is the masked, area-weighted global average over valid grid cells for each channel. The anomaly RMSE is

$$\mathrm{RMSE}(t) = \sqrt{\left\langle (\bar{x}_t - y_t)^2 \right\rangle_w}. \tag{5}$$

Probabilistic skill is evaluated using fair CRPS, also in anomaly space:

$$\mathrm{CRPS}(t) = \left\langle \frac{1}{E}\sum_{i=1}^{E} |x_{i,t} - y_t| - \frac{1}{E(E-1)}\sum_{i<j} |x_{i,t} - x_{j,t}| \right\rangle_w. \tag{6}$$

Ensemble calibration is summarized with the spread–skill ratio (SSR). The ensemble spread is computed from the unbiased ensemble variance,

$$\mathrm{Spread}(t) = \left[\left\langle \frac{1}{E-1}\sum_{i=1}^{E} (x_{i,t} - \bar{x}_t)^2 \right\rangle_w\right]^{1/2}, \tag{7}$$

and

$$\mathrm{SSR}(t) = \frac{\mathrm{Spread}(t)}{\mathrm{RMSE}(t)}. \tag{8}$$

Values near one indicate that the ensemble spread is consistent with the ensemble-mean error [67]; values below one indicate underdispersion and values above one indicate overdispersion.

For the correlation metrics, we additionally remove an observed warming trend estimated from ERA5, in line with standard evaluation protocols to isolate seasonal skill [13, 16]. For each calendar month, channel, and grid point, we fit a linear model to ERA5 monthly means over 1990–2019. For target time t, the trend-aware climatology is the fitted value

$$C_t^{\mathrm{tr}} = \hat{a}_{m(t)} + \hat{b}_{m(t)}\,\tau(t), \tag{9}$$

where $m(t)$ and $\tau(t)$ are the calendar month and year of the target time. For the trend-aware climatology, a reference period of 1990–2019 is chosen to overlap with training years. The fitted coefficients are specific to each month, channel, and grid point. The detrended anomalies are then formed by subtracting C^{tr} instead of the stationary climatology:

$$x_{i,t}^{\mathrm{dt}} = X_{i,t}^{\mathrm{db}} - C_t^{\mathrm{tr}}, \qquad y_t^{\mathrm{dt}} = Y_t - C_t^{\mathrm{tr}}, \qquad \bar{x}_t^{\mathrm{dt}} = \frac{1}{E}\sum_{i=1}^{E} x_{i,t}^{\mathrm{dt}}. \tag{10}$$

The reported anomaly correlation coefficient (ACC) is computed from these detrended anomalies:

$$\mathrm{ACC}(t) = \frac{\left\langle \bar{x}_t^{\mathrm{dt}} y_t^{\mathrm{dt}} \right\rangle_w}{\sqrt{\left\langle (\bar{x}_t^{\mathrm{dt}})^2 \right\rangle_w \left\langle (y_t^{\mathrm{dt}})^2 \right\rangle_w}}. \tag{11}$$

The temporal anomaly correlation coefficient (TCC) is computed at each grid point by correlating detrended anomalies over initialization dates for a given lead time:

$$\mathrm{TCC} = \left\langle \frac{\sum_t \bar{x}_t^{\mathrm{dt}} y_t^{\mathrm{dt}}}{\sqrt{\sum_t \left(\bar{x}_t^{\mathrm{dt}}\right)^2 \sum_t \left(y_t^{\mathrm{dt}}\right)^2}} \right\rangle_w . \tag{12}$$

For the trend-aware skill scores, the reference forecast is the ERA5 linear-trend climatology. Let

$$\delta_t = C_t^{\mathrm{tr}} - C_t \tag{13}$$

denote the climatological anomaly. The deterministic trend-climatology forecast in anomaly space is therefore δ_t, giving

$$\mathrm{RMSE}_{\text{tr-clim}}(t) = \sqrt{\left\langle \left(\delta_t - y_t\right)^2 \right\rangle_w}. \tag{14}$$

For CRPSS, the reference is a trend-aware climatological ensemble. For each training year j and calendar month m, we compute the ERA5 residual about the fitted trend,

$$r_{j,m} = Z_{j,m} - C_{j,m}^{\mathrm{tr}}, \tag{15}$$

where $Z_{j,m}$ denotes the ERA5 field in historical year j and calendar month m and $C^{\mathrm{tr}}_{j,m}$ is the fitted trend climatology. For target time t, with calendar month $m(t)$, the anomaly-space climatological ensemble members are

$$z^{\text{tr-clim}}_{j,t} = \delta_t + r_{j,m(t)}. \tag{16}$$

The denominator $\mathrm{CRPS}_{\text{tr-clim}}(t)$ is the fair CRPS of this climatological ensemble against y_t. The reported skill scores are

$$\mathrm{CRPSS}(t) = 1 - \frac{\mathrm{CRPS}(t)}{\mathrm{CRPS}_{\text{tr-clim}}(t)}, \qquad \mathrm{RMSS}(t) = 1 - \frac{\mathrm{RMSE}(t)}{\mathrm{RMSE}_{\text{tr-clim}}(t)}. \tag{17}$$

Positive values indicate improvement over the trend-aware climatological baseline. Statistical significance is assessed using 1000 bootstrap resamples of the initialization dates. In each resample, we randomly draw the same number of dates as in the original evaluation set, with replacement. We then recompute the mean quantity for each resample. A result is considered significant when the central 95% of these 1000 bootstrap estimates does not include zero.

To summarize how much skill FGT retains relative to SEAS5, we compute a skill-retention ratio between their initialization-averaged scores. Let S^A and S^B denote the scores of model A and baseline model B, respectively, after averaging over initialization dates. For positively oriented metrics such as RMSS, CRPSS, ACC, and TCC, we define retention as

$$R^{A/B} = \begin{cases} S^A/S^B, & S^A > 0,\ S^B > 0, \\ 0, & S^A \le 0,\ S^B > S^A, \\ 2, & S^A > S^B,\ S^B \le 0, \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{18}$$

For SSR, we first convert the raw spread–skill ratio to a calibration score,

$$S_{\mathrm{SSR}} = \exp\left(-\left|\log \mathrm{SSR}\right|\right), \tag{19}$$

and then apply the same retention definition in Eq. (18). This score is maximized when $\mathrm{SSR} = 1$, so $R^{A/B}_{\mathrm{SSR}} > 1$ means that model A is closer than model B to the ideal spread–skill ratio.

4.4 Computation of seasonal indices

The large-scale indices are computed from the same anomaly fields used in Sec. 4.3. These indices use the stationary climatologies M_t and C_t. Using the sea-surface temperature anomaly, the monthly Niño 3.4 anomaly for ensemble member i is

$$\text{Niño } 3.4'_{i,t} = \langle x_{i,t} \rangle_{w,\ 5^\circ\mathrm{S}-5^\circ\mathrm{N},\ 170^\circ\mathrm{W}-120^\circ\mathrm{W}}, \tag{20}$$

and the monthly DMI anomaly is

$$\mathrm{DMI}'_{i,t} = \langle x_{i,t} \rangle_{w,\ 10^\circ\mathrm{S}-10^\circ\mathrm{N},\ 50^\circ\mathrm{E}-70^\circ\mathrm{E}} - \langle x_{i,t} \rangle_{w,\ 10^\circ\mathrm{S}-0^\circ,\ 90^\circ\mathrm{E}-110^\circ\mathrm{E}}\,. \tag{21}$$

The NAO index is computed using mean sea-level pressure. We project monthly anomalies onto a fixed leading empirical orthogonal function (EOF1) pattern from ERA5. The EOF e_1 is fitted once from ERA5 DJF monthly mean sea-level pressure anomalies over 20°N–80°N, 90°W–40°E, over 1993–2016. Its sign is chosen so that the Azores-minus-Iceland loading is positive. For DJF target months,

$$\mathrm{NAO}'_{i,t} = \frac{1}{\sigma_{\mathrm{PC1}}} \sum_{r \in R_{\mathrm{NAO}}} x_{i,t}(r) \sqrt{\cos \phi_r}\, e_1(r), \tag{22}$$

where $r = (\phi_r, \lambda_r)$ denotes latitude and longitude of the EOF region R_{NAO}, and σ_{PC1} is the standard deviation of the first ERA5 principal component (PC1) over the fitting period.

ENSO teleconnection maps show, for each field separately, the gridpoint regression of monthly field anomalies onto Niño 3.4 anomalies. At each grid point r and lead month ℓ, the response variable is the anomaly of the plotted field, for example, precipitation. The plotted coefficient is

$$\beta_\ell(r) = \frac{\sum_s (x_s - \bar{x})\,(y_{s,\ell}(r) - \bar{y}_\ell(r))}{\sum_s (x_s - \bar{x})^2}, \tag{23}$$

where s indexes the samples used in the regression, x_s is the Niño 3.4 anomaly for sample s, and $y_{s,\ell}(r)$ is the corresponding field anomaly at lead month ℓ and grid point r. The quantities $\bar{x}$ and $\bar{y}_\ell(r)$ are the corresponding sample means over s and are subtracted to center the regression data.

4.5 Use of generative AI tools

OpenAI Codex was used in software development to write portions of the codebase and to assist with refactoring and debugging. All AI-assisted code was reviewed, modified as necessary, and tested by the authors, who take full responsibility for the final implementation and the reported results.

Supplementary information. Supplementary information containing additional results accompanies this manuscript. In the preprint version, the Supplementary Information is appended directly to the main text.

Acknowledgements. SA was funded by the Helmholtz Foundation Model Initiative through the project *Representation Model for Climate Science (HClimRep)*. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project IDs bg1551 and gg0028.

5 Data availability

Precomputed FGT forecasts and climatologies supporting the analyses are publicly available from the Hugging Face repository [29].

The ERA5 fields used for training and verification are publicly available from the WeatherBench2 archive and, for the extended period not covered by WeatherBench2, from the Copernicus Climate Data Store. The greenhouse-gas and stratospheric aerosol forcing data are publicly available from the CMIP6 input4MIPs and NASA GISS datasets. The SEAS5 reference consists of 51-member real-time forecasts initialized on the first day of each month using the fixed IFS Cycle 43r1 model configuration; these forecasts and the AI model forecasts used for comparison are available from public archives [5, 28, 68].

6 Code availability

Code for data preprocessing, training, inference, and forecast evaluation will be released in a public repository after peer review is complete. For peer review, a private copy of the code is made available to the editors and reviewers via Helmholtz GitLab.

7 Author contributions

Conceptualization: SA, EZ, DG;
Methodology: SA, EZ;
Software: SA;
Validation: SA;
Formal analysis: SA;
Investigation: SA;
Resources: DG, EZ;
Data curation: SA;
Writing – original draft: SA, AB, AF;
Writing – review & editing: SA, AB, AF, EZ, DG;
Visualization: SA;
Supervision: DG;
Project administration: DG;
Funding acquisition: DG.

Competing interests. The authors declare no competing interests.

References

[1] Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P. & Rodrigues, L. R. L. Seasonal climate predictability and forecasting: status and prospects. *WIREs Climate Change* **4**, 245–268 (2013).

[2] White, C. J., Domeisen, D. I. V., Acharya, N. *et al.* Advances in the application and utility of subseasonal-to-seasonal predictions. *Bulletin of the American Meteorological Society* **103**, E1448–E1472 (2022).

[3] Barriopedro, D., García-Herrera, R., Ordóñez, C., Miralles, D. G. & Salcedo-Sanz, S. Heat waves: Physical understanding and scientific challenges. *Reviews of Geophysics* **61**, e2022RG000780 (2023).

[4] Keane, R. J., Parker, D. J., Dunn-Sigouin, E., Kolstad, E. W. & Marsham, J. H. Mid-latitude versus tropical scales of predictability and their implications for forecasting. *Meteorological Applications* **32**, e70055 (2025).

[5] Johnson, S. J. *et al.* SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development* **12**, 1087–1117 (2019).

[6] Monhart, S. *et al.* Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres* **123**, 7999–8016 (2018).

[7] Mouatadid, S. *et al.* Adaptive bias correction for improved subseasonal forecasting. *Nature Communications* **14**, 3482 (2023).

[8] Pathak, J. *et al.* FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214 (2022). URL https://arxiv.org/abs/2202.11214.

[9] Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).

[10] Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).

[11] Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).

[12] Shukla, P. & Halem, M. DUNE: A machine learning deep U-Net++-based intrinsic forecast approach to monthly, seasonal, and annual climate forecasting. *Artificial Intelligence for the Earth Systems* **4** (2025).

[13] Palma, L. *et al.* Data-driven seasonal climate predictions via variational inference and transformers. *npj Climate and Atmospheric Science* **9**, 48 (2026).

[14] Yuan, Y., Ding, J., Qiu, Z., Fan, J. & Li, Y. Learning the coupled dynamics of global climate modes. *Nature Machine Intelligence* **8**, 930–941 (2026).

[15] Yu, Y. *et al.* ClimateAR: Multi-scale autoregressive generative modeling for seasonal-to-interannual climate forecasting. International Conference on Machine Learning (ICML) (2026).

[16] Chen, L. *et al.* A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications* **15**, 6425 (2024).

[17] Li, G. *et al.* TianQuan-S2S: A subseasonal-to-seasonal global weather model via incorporate climatology state. International Conference on Learning Representations (ICLR) (2026). URL https://openreview.net/forum?id=7Dvmq7MhwU.

[18] Schloer, J. *et al.* AIFS-SUBS: Extending data-driven forecasting to sub-seasonal timescales. arXiv preprint arXiv:2607.05100 (2026).

[19] Kent, C. *et al.* Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data. *npj Climate and Atmospheric Science* **8**, 314 (2025).

[20] Antonio, B., Strommen, K. & Christensen, H. M. Seasonal forecasting using the GenCast probabilistic machine learning model. *Climate Dynamics* **64**, 148 (2026).

[21] Wang, C. *et al.* Coupled ocean–atmosphere dynamics in a machine learning earth system model. arXiv preprint arXiv:2406.08632 (2024).

[22] Cresswell-Clay, N. *et al.* A deep learning earth system model for efficient simulation of the observed climate. *AGU Advances* **6**, e2025AV001706 (2025).

[23] Wu, H. *et al.* Advanced long-term earth system forecasting. arXiv preprint arXiv:2505.19432 (2025). Version 3, revised 9 January 2026.

[24] Ferro, C. A. T., Richardson, D. S. & Weigel, A. P. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications* **15**, 19–24 (2008).

[25] Camps-Valls, G. *et al.* Bridging the weather and climate divide with artificial intelligence. *Nature Communications* **17**, 8578 (2026).

[26] Alet, F. *et al.* Skillful joint probabilistic weather forecasting from marginals. arXiv preprint arXiv:2506.10772 (2025).

[27] Vitart, F. Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society* **140**, 1889–1899 (2014).

[28] Rasp, S. *et al.* WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems* **16**, e2023MS004019 (2024).

[29] Agarwal, S. FGT. Hugging Face dataset repository (2026). URL https://doi.org/10.57967/hf/9775.

[30] Lehmann, F. *et al.* Can AI weather models predict beyond two weeks? a quantitative benchmark and analysis of long rollouts. arXiv preprint arXiv:2605.30184

(2026).

[31] Clark, S. K. *et al.* Disentangling the effects of sea surface temperature and CO_2 in global machine learned weather-climate emulators. *arXiv preprint arXiv:2606.07928* (2026).

[32] Wong, J. *et al.* D5.1: Definition of Demonstrator Applications and Workflows (2023). URL https://horizon-opencube.eu/wp-content/uploads/2024/06/D5.1-Definition-of-Demonstrator-Applications-and-Workflows.pdf. Horizon Europe Project 101092984, final version 2023-12-21.

[33] Ling, F. *et al.* Multi-task machine learning improves multi-seasonal prediction of the Indian Ocean Dipole. *Nature Communications* **13**, 7681 (2022).

[34] Silverman, B. W. *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).

[35] Alet, F. *et al.* Operational tropical cyclone forecasting with ai. *Nature* (2026).

[36] Hahner, S. *et al.* Representing the surface ocean in ECMWF's data-driven forecasting system AIFS. arXiv preprint arXiv:2604.25559 (2026).

[37] Boucher, E. *et al.* Learning coupled earth system dynamics with GraphDOP. arXiv preprint arXiv:2510.20416 (2025).

[38] Duncan, J. P. C. *et al.* SamudrACE: Fast and accurate coupled climate modeling with 3D ocean and atmosphere emulators. arXiv preprint arXiv:2509.12490 (2025).

[39] Sausen, R., Barthel, K. & Hasselmann, K. Coupled ocean-atmosphere models with flux correction. *Climate Dynamics* **2**, 145–163 (1988).

[40] Agarwal, S., Tosi, N., Kessel, P., Breuer, D. & Montavon, G. Deep learning for surrogate modeling of two-dimensional mantle convection. *Physical Review Fluids* **6**, 113801 (2021).

[41] Bodnar, C. *et al.* A foundation model for the Earth system. *Nature* **641**, 1180–1187 (2025).

[42] Hersbach, H. *et al.* The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049 (2020).

[43] Copernicus Climate Change Service. ERA5 hourly data on single levels from 1940 to present (2018).

[44] Copernicus Climate Change Service. ERA5 hourly data on pressure levels from 1940 to present (2018).

[45] Meinshausen, M. *et al.* Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geoscientific Model Development* **10**, 2057–2116 (2017).

[46] NASA Goddard Institute for Space Studies. Original GISS stratospheric aerosol data. https://data.giss.nasa.gov/modelforce/strataer/original_GISS_data/ (2016). Accessed 17 June 2026.

[47] Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lecture Notes in Computer Science, 234–241 (2015).

[48] Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B. & Wetzstein, G. Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (2020). URL https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html.

[49] Bekar, A. C., Agarwal, S., Hüttig, C., Tosi, N. & Greenberg, D. S. Hybrid latent representations for PDE emulation. Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025) (2025). URL https://openreview.net/forum?id=Hh8ebJYQs3.

[50] Lang, S. *et al.* AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence* **2**, 18 (2026).

[51] Bonev, B. *et al.* FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. arXiv preprint arXiv:2507.12144 (2025).

[52] Zhdanov, M., Lucic, A., Welling, M. & van de Meent, J.-W. (sparse) attention to the details: Preserving spectral fidelity in ML-based weather forecasting models. Forty-third International Conference on Machine Learning (ICML 2026) (2026). URL https://openreview.net/forum?id=Trfut9gSjY.

[53] Andry, G. *et al.* Appa: Bending weather dynamics with latent diffusion models for global data assimilation. Machine Learning and the Physical Sciences Workshop at NeurIPS 2025 (2025).

[54] Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**, 3942–3951 (2018).

[55] Peebles, W. & Xie, S. Scalable diffusion models with transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 4195–4205 (2023).

[56] Agarwal, S., Bekar, A. C., Hüttig, C., Greenberg, D. S. & Tosi, N. Physics-based machine learning for mantle convection simulations. *Physics of Fluids* **37**, 086624 (2025).

[57] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007).

[58] Ferro, C. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society* **140**, 1917–1923 (2014).

[59] Zamo, M. & Naveau, P. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences* **50**, 209–234 (2018).

[60] Brandstetter, J., Worrall, D. E. & Welling, M. Message passing neural PDE solvers. International Conference on Learning Representations (ICLR) (2022). URL https://openreview.net/forum?id=vSix3HPYKSU.

[61] Krouma, M., Yiou, P., Déandreis, C. & Thao, S. Assessment of stochastic weather forecast of precipitation near European cities, based on analogs of circulation. *Geoscientific Model Development* **15**, 4941–4958 (2022).

[62] Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B. & Searight, K. Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review* **141**, 3498–3516 (2013).

[63] Tokuda, D. & Dirmeyer, P. A. Selective reuse of prior ensemble data improves the latest air temperature forecast over North America. *Proceedings of the National Academy of Sciences* **123**, e2524516123 (2026).

[64] Carrassi, A., Bocquet, M., Bertino, L. & Evensen, G. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change* **9**, e535 (2018).

[65] Batté, L. *et al.* Summer predictions of Arctic sea ice edge in multi-model seasonal re-forecasts. *Climate Dynamics* **54**, 5013–5029 (2020).

[66] Zampieri, L., Goessling, H. F. & Jung, T. Bright prospects for Arctic sea ice prediction on subseasonal time scales. *Geophysical Research Letters* **45**, 9731–9738 (2018).

[67] Fortin, V., Abaza, M., Anctil, F. & Turcotte, R. Why should ensemble spread match the RMSE of the ensemble mean? *Journal of Hydrometeorology* **15**, 1708–1713 (2014).

[68] Zhong, X. FuXi-S2S (2025).

[69] Vyas, N. *et al.* Soap: Improving and stabilizing shampoo using adam (2025). URL https://arxiv.org/abs/2409.11321. arXiv:2409.11321.

[70] TritonCast. Tritoncast inference datasets: Atmospheric year-long forecast. Hugging Face dataset repository (2025). URL https://huggingface.co/datasets/TritonCast/TritonCast_inference_datasets/tree/19f28f9f827b52a359312cae56e8bbf6a5334a07/Atmospheric_year_long_forecast.

Supplementary Information

Learning skillful forecasting from days to seasons

Siddhant Agarwal, Ali Bekar, Anthony Frion, Eduardo Zorita, and David Greenberg

Contents

1 FGT implementation details **29**
- 1.1 ERA5 dataset used to train FGT . . . 29
- 1.2 Architecture hyperparameters . . . 29
- 1.3 Self- and pooling-attention blocks . . . 30
- 1.4 Training curriculum . . . 31
- 1.5 Inference-time ensemble pruning . . . 31

2 Ablation experiments **32**

3 Prediction of the February 2021 USA cold wave from a medium-range initialization **33**

4 Multihorizon skill evaluation **33**
- 4.1 Evaluation protocol for medium-range and subseasonal forecasts . . . 33
- 4.2 Comparison with recent AI studies at seasonal lead times . . . 34
- 4.3 Long-range prediction: stability versus skill . . . 37

5 Seasonal skill of FGT **39**
- 5.1 ENSO precipitation teleconnections . . . 39
- 5.2 Spatial distribution of skill . . . 40
- 5.3 Global skill evaluation . . . 40
- 5.4 FGT performance on additional unseen years . . . 40

6 Prediction snapshots **41**

Table S1 ERA5 variables used for training and evaluation. The Input and Output columns indicate whether each variable is provided to the network as an autoregressive input and whether it is predicted by the network. Precipitation is predicted diagnostically but is not fed back as an input. Top-of-atmosphere downward shortwave radiation flux is analytically prescribed and is therefore used as an input but not predicted.

Variable	Display name	Input	Output
2-m temperature [K]	tas	Yes	Yes
10-m zonal wind [m s^{-1}]	uas	Yes	Yes
10-m meridional wind [m s^{-1}]	vas	Yes	Yes
Mean sea-level pressure [Pa]	msl	Yes	Yes
Total precipitation [mm day^{-1}]	pr	No	Yes
Sea-surface temperature [K]	tos	Yes	Yes
Sea-ice concentration [1]	siconc	Yes	Yes
Total soil moisture content [kg m^{-2}]	mrso	Yes	Yes
Surface net longwave radiation flux [W m^{-2}]	sfc net longwave	Yes	Yes
Surface net shortwave radiation flux [W m^{-2}]	sfc net shortwave	Yes	Yes
TOA downward shortwave radiation flux [W m^{-2}]	TOA shortwave	Yes	No
TOA net longwave radiation flux [W m^{-2}]	TOA net longwave	Yes	Yes
Atmospheric variables at 1000, 925, 850, 700, 500, 300, 200, 100, and 10 hPa			
Zonal wind [m s^{-1}]	ua	Yes	Yes
Meridional wind [m s^{-1}]	va	Yes	Yes
Geopotential height [m]	zg	Yes	Yes
Air temperature [K]	ta	Yes	Yes
Specific humidity [kg kg^{-1}]	hus	Yes	Yes
Static spatial fields			
Orography	orog	Yes	No
Spherical-coordinate SIREN embedding	coord emb	Yes	No
Land-sea mask	mask	Yes	No

1 FGT implementation details

1.1 ERA5 dataset used to train FGT

Table S1 summarizes the input and output variables for the Functional Graph Transformer (FGT).

1.2 Architecture hyperparameters

Table S2 lists the hyperparameters for FGT described in Sec. 4.2.1. Given the large training cost associated with such models, hyperparameters were selected manually during preliminary experiments on UFS-Replay to balance model capacity, optimization stability, GPU-memory requirements, and the feasibility of long autoregressive rollouts. The resulting 35-million-parameter architecture was fixed before training on the ERA5 dataset.

Table S2 FGT hyperparameters.

Component	Value
Icosahedral nodes at levels	$40962, 10242, 2562, 642$
Node feature width	256 channels at every graph level
Attention heads	4 at every graph level
Attention dimensions	$d_q = d_k = d_v = 64$ per head
MLP hidden width	1024, with no dropout
Encoder self-attention blocks	$3, 3, 3, 3$ from fine to coarse graph levels
Decoder self-attention blocks	$3, 3, 3, 3$ from coarse to fine graph levels
Encoder pooling attention operations	1 per level
Decoder unpooling attention operations	1 per level
Self-attention neighborhoods	$3, 3, 3, 8$ graph hops from fine to coarse levels
Pooling neighborhoods	3-hop neighborhoods
Activation	SiLU
Normalization	LayerNorm without affine parameters
Noise dimension	$z_i \in \mathbb{R}^{64}$
Noise embedding MLP	2 hidden layers with 64 features each
Noise embedding scale	Trainable scalar initialized to 10^{-3}
Optimizer	SOAP [69]
Batch size	64
Gradient clipping	1.0
Weight decay	0.01

1.3 Self- and pooling-attention blocks

We use the sparse icosahedral attention computation introduced by Appa, but modify the surrounding blocks for autoregressive stability. Self-attention and MLP updates are applied sequentially through independently gated residual branches, while pooling includes a direct source-to-target skip connection to propagate local information.

$$\tilde{x} = \mathrm{AdaLN}_a(x, c_i), \tag{1}$$

$$y = x + g_a(c_i) \odot \mathrm{Attn}(q = \tilde{x}, k = \tilde{x}, v = \tilde{x}), \tag{2}$$

$$\tilde{y} = \mathrm{AdaLN}_m(y, c_i), \tag{3}$$

$$\mathrm{SelfAttn}(x) = y + g_m(c_i) \odot \mathrm{MLP}(\tilde{y}). \tag{4}$$

Here c_i is the noise-modulation embedding for ensemble member i, and $g_a(c_i)$ and $g_m(c_i)$ are tanh-bounded learned gates applied to the attention and MLP operations, respectively. We first embed z_i as $\hat{z}_i = s_z E_z(z_i) \in \mathbb{R}^{64}$ and then set $c_i = E_{\mathrm{mod}}([\Delta t, \hat{z}_i]) \in \mathbb{R}^{128}$, where E_z is the noise-embedding MLP, s_z is its trainable scale, and E_{mod} is the block-conditioning MLP. Including Δt allows the architecture to support heterogeneous time increments. All models used in this study, however, operate at a fixed one-day increment, so $\Delta t = 1$ does not vary between training examples and acts only as a constant conditioning input in the present experiments.

Pooling is performed by sparse cross-attention from the source nodes to the target nodes. Let x_q denote the closest-source-node features associated with the target

Table S3 FGT autoregressive training stages. Learning-rate schedules use optimizer steps.

Stage	Steps	Rollout horizon	Ensemble size	Learning-rate schedule	Warmup
1	5500	Curriculum $1 \to 8$	4	Cosine, $10^{-3} \to 10^{-5}$	200
2	1000	15	4	Cosine, $10^{-4} \to 10^{-5}$	100
3	1000	Mixture of 9, 15, 30	2,4,6	Cosine, $10^{-5} \to 10^{-6}$	50
4	950	Curriculum $30 \to 180$	2,4,6	Constant, 10^{-6}	50

Table S4 Rollout–ensemble sampling and scoring windows in stages 3 and 4.

Stage	Rollout:Ensemble	Sampling	Scored rollout steps
3	30:2, 15:4, 9:6	0.33, 0.34, 0.33	All steps in the sampled rollout.
4	30:2, 15:4, 9:6	0.33, 0.34, 0.33	3 equal windows in the 180-step rollout: initial, random middle, and tail

queries, and let x_k denote the source-node features.

$$q_f = \mathrm{LN}(x_q + \phi(r^q)), \tag{5}$$

$$k_f = \mathrm{LN}\big(x_k + \phi(r^k)\big), \tag{6}$$

$$a = \mathrm{Attn}(q = q_f, k = k_f, v = k_f) + b(r^q), \tag{7}$$

$$y = W_{\mathrm{skip}} x_q + g_a(c_i)\, a, \tag{8}$$

$$\tilde{y} = \mathrm{AdaLN}_m(y, c_i), \tag{9}$$

$$\mathrm{PoolAttn}(x) = W_{\mathrm{out}}[y + g_m(c_i)\, \mathrm{MLP}(\tilde{y})]. \tag{10}$$

Here $r^q, r^k \in \mathbb{R}^3$ are the absolute unit-sphere Cartesian coordinates of the target and source nodes, respectively. A shared SIREN ϕ embeds the two coordinate sets independently, while a separate SIREN $b(r^q)$ provides an additive target-coordinate feature after cross-attention.

The learned gates $g_a(c_i)$ and $g_m(c_i)$ are tanh-bounded and broadcast over nodes and channels. The pooling attention itself is not modulated by c_i; conditioning is applied only through the two gates and the AdaLN in the MLP branch. W_{skip} and W_{out} are linear projections. Gate parameters are separate for every self-attention and pooling block.

1.4 Training curriculum

FGT is trained in four stages with the settings shown in Table S3 and in Table S4.

1.5 Inference-time ensemble pruning

At inference, candidate ensemble trajectories are generated by sampling independent Gaussian noise vectors. To select an ensemble using only information available at forecast initialization, we run a two-day pre-initialization spin-up. Candidate members

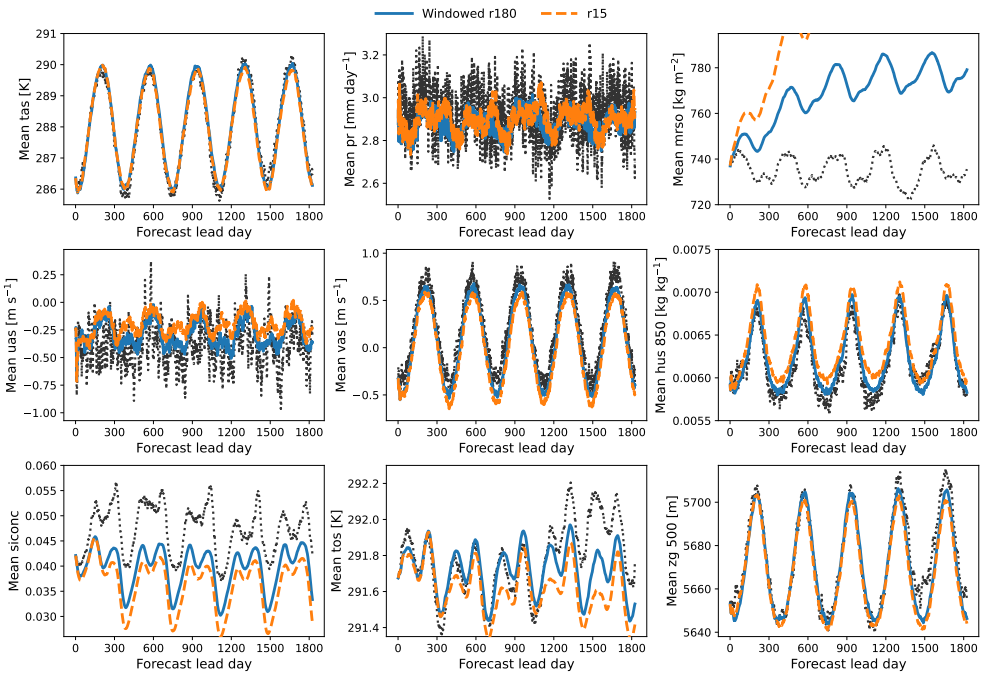


Fig. S1 Global-mean trajectories of selected atmospheric and land variables during a five-year forecast. The orange dashed line denotes the Short rollout configuration ($r15$), the blue line denotes the Reference configuration (the window-trained $r180$ model), and the black dotted line denotes the corresponding ERA5 seasonal evolution. The reduced drift of the Reference configuration indicates that long-rollout training limits the accumulation of forecast errors. All trajectories show raw FGT predictions without bias correction.

are initialized two days before the forecast start and scored against the observed states at $t_0 - 1$ and t_0. The selected members are then continued from t_0 to the forecast target. No information from times after t_0, including the verification target, is used in member selection.

Member selection is done as follows. For each candidate member, we compute cosine-latitude-weighted global means for the following channels: zg 500, zg 850, uas, vas, hus 850, tas, tos, mrso, pr. These variables were chosen to summarize the large-scale atmospheric circulation, near-surface thermodynamic state, ocean and land-surface boundary conditions, and atmospheric and land moisture. The score is averaged over the selected variables and over the two spin-up verification times, and the members with the lowest RMSE are retained.

2 Ablation experiments

We compare the Reference configuration with five ablations designed to isolate the effects of rollout horizon, the temporal placement of the training loss, slow-surface evolution, and inference-time ensemble pruning. The compute-matched Short rollout

configuration retains a maximum training rollout of 15 steps with the same compute budget. The Last window configuration computes the training loss in stage 4 only over one contiguous window instead of the three different scoring windows at the start, a random intermediate point, and the end. The Persisted t/s/m experiment holds sea-surface-temperature, sea-ice-concentration, and soil-moisture anomalies fixed at their initialized values throughout inference. The No selection experiment disables inference-time pruning by retaining all 51 generated members rather than selecting 51 members from 300 candidates. Finally, the No t/s/m experiment excludes sea-surface temperature, sea-ice concentration, and soil moisture from the predicted state during training and inference.

Fig. S1 illustrates the effect of training-rollout length on long-term stability. Over five-year predictions, the Short rollout configuration develops greater drift in several variables, whereas the Reference configuration more closely follows the seasonal evolution represented by ERA5.

Table S5 and Fig. 3 report debiased test-set performance for the six configurations. The comparison between the Short rollout and Reference configurations isolates the effect of extending the training rollout. The Last window experiment further tests whether long-rollout performance depends only on supervising the end of the rollout or benefits from distributing the loss across different forecast stages. The Persisted t/s/m experiment quantifies the contribution of dynamically evolving slow-surface variables, while the No selection experiment measures the benefit of inference-time candidate selection. The No t/s/m experiment tests how including these coupled surface variables affects skill in the remaining predicted fields; scores for the omitted variables are undefined and are shown as dashes in Table S5.

3 Prediction of the February 2021 USA cold wave from a medium-range initialization

Forecasts associated with the February 2021 USA cold wave are displayed in Fig. S2. FGT captures the broad cold anomaly, although it underestimates its magnitude and places the strongest cooling too far north. Its 2-m temperature ACC is 0.61. SEAS5 instead produces warm anomalies over much of the western and central United States and has little spatial agreement with ERA5 (ACC = -0.08). FGT represents more of the broad negative 500-hPa geopotential-height anomaly over central and eastern North America, whereas SEAS5 produces a less correlated height pattern. The geopotential-height agreement remains modest.

4 Multihorizon skill evaluation

4.1 Evaluation protocol for medium-range and subseasonal forecasts

The evaluation protocol is summarized in Table S6. Forecast skill was evaluated against ERA5 using only completed initialization dates available to both models in each comparison. Daily and weekly forecasts were evaluated on the common 1.5° grid, whereas

Table S5 Ablation summary for selected variables. CRPSS values are averaged over lead months 1–6 for the six debiased FGT configurations. RMSS is additionally shown for sea-surface temperature, sea ice, and soil moisture, as these fields have no ensemble spread by construction when their anomalies are persisted in the Persisted t/s/m experiment. Bold values indicate the strongest statistically significant configurations within each row.

Variable	Reference	Persisted t/s/m	No t/s/m	Short rollout	Last window	No selection
CRPSS						
tas	0.067	0.048	**0.081**	-0.023	0.035	0.051
uas	0.092	0.075	**0.096**	0.087	0.091	0.070
vas	**0.088**	0.068	0.084	**0.086**	**0.088**	0.052
msl	**0.098**	0.086	**0.099**	0.085	**0.096**	0.082
pr	**0.098**	0.056	**0.099**	**0.096**	0.094	0.037
sfc long wave	**0.082**	0.061	**0.081**	0.069	0.072	0.017
sfc short wave	**0.079**	0.041	**0.080**	**0.077**	0.074	-0.059
top long wave	**0.104**	0.058	**0.101**	0.090	0.095	0.044
ua 850	**0.096**	0.079	**0.099**	0.089	**0.095**	0.080
ua 500	**0.095**	0.084	**0.093**	0.091	**0.094**	0.082
ua 200	**0.114**	0.089	**0.113**	0.109	0.106	0.099
va 850	**0.080**	0.068	**0.078**	**0.080**	**0.079**	0.041
va 500	**0.078**	0.071	**0.076**	**0.077**	**0.076**	0.047
va 200	**0.086**	0.069	**0.085**	**0.084**	**0.084**	0.072
zg 850	0.099	0.094	**0.107**	0.098	0.096	0.086
zg 500	0.091	**0.122**	0.113	0.049	0.074	0.089
zg 200	0.091	**0.156**	0.123	-0.027	0.053	0.088
ta 850	0.076	**0.105**	0.093	0.012	0.051	0.078
ta 500	0.057	**0.115**	0.093	-0.073	0.011	0.064
ta 200	0.110	**0.135**	**0.138**	0.041	0.083	0.092
hus 850	0.051	**0.063**	**0.064**	-0.022	0.020	-0.001
hus 500	**0.099**	0.062	**0.098**	0.052	0.082	0.041
hus 200	**0.085**	**0.091**	0.077	-0.121	0.037	0.017
RMSS						
tos	**0.132**	-0.093	–	**0.117**	**0.117**	**0.119**
siconc	0.116	-0.080	–	0.116	**0.126**	-0.032
mrso	**0.505**	0.476	–	0.497	0.496	**0.503**

monthly forecasts were evaluated on the 1° ERA5 grid. Forecast debiasing was distinct from the trend-aware climatology used to calculate skill. Daily forecasts were evaluated without forecast debiasing, weekly forecasts were debiased using model and ERA5 climatologies from 2002–2016, and monthly forecasts used climatologies from 1993–2016. At every horizon, the RMSS and CRPSS climatological reference was derived from the 1990–2019 ERA5 linear-trend climatology and evaluated at the forecast valid time.

4.2 Comparison with recent AI studies at seasonal lead times

Recent AI studies extend forecasting to seasonal lead times, but differ in their forecast targets, temporal aggregation, treatment of surface boundary conditions, and publicly available outputs. Table S8 summarizes the main differences relevant to the monthly comparison in Fig. 2. Forecast archives provided by the authors of these studies would be the most reliable way to avoid any inference errors. Otherwise, the following are

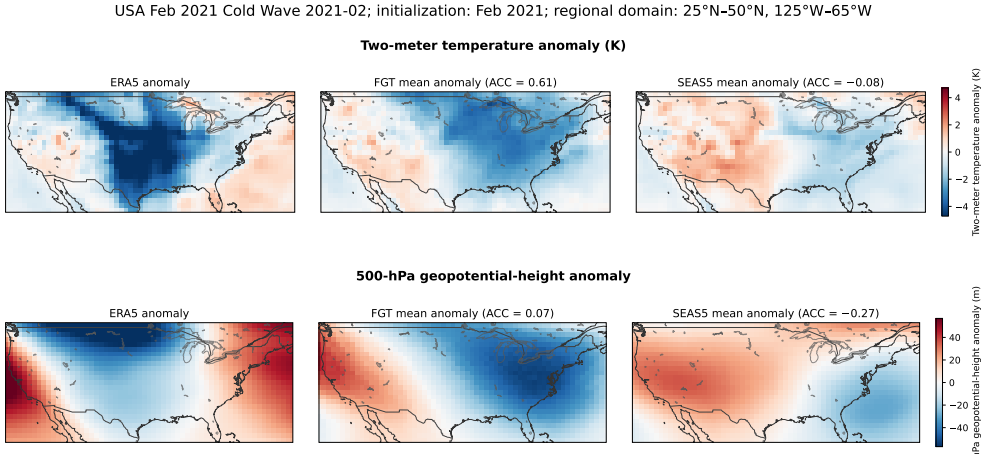


Fig. S2 Medium-range forecast of the February 2021 USA cold wave, initialized on 1 February 2021. The upper row shows 2-m temperature anomalies, and the lower row shows 500-hPa geopotential-height anomalies. Columns show the ERA5 anomaly and the FGT and ECMWF SEAS5 ensemble-mean anomalies. Values in parentheses give the detrended spatial ACC against ERA5 over the displayed region.

Table S6 Evaluation, ensemble size, forecast-debiasing climatology, and trend-aware ERA5 climatology. Evaluation periods refer to common completed initialization dates. Ensemble sizes are reported as FGT/reference.

Comparison	Evaluation	Resolution	Members	Debiasing	Climatology
FGT–WN2	2022	1.5°	51/51	Not applied	1990–2019
FGT–GenCast	2020	1.5°	51/51	Not applied	1990–2019
FGT–FuXi-S2S	2020–2021	1.5°	51/51	2002–2016	1990–2019
FGT–IFS	2020–2022	1.5°	51/50	2002–2016	1990–2019
FGT–SEAS5	2020–2026	1°	51/51	1993–2016	1990–2019

Table S7 Approximate model size, computational cost, resolution, and inference speed of selected global neural forecasting models. Values are reported for the published or publicly released configurations.

Model	Resolution	Step	Fields	Parameters	Training cost	Inference [s]
This work	1°	1 d	56	35M	115 GPU-days	0.11
WN2	0.25°	6 h	84	4×180M	4×490 TPU-days	< 4.0
GenCast	0.25°	12 h	84	57M	∼160 TPU-days	32
FuXi-S2S	1.5°	1 d	76	n.r.	17k steps on 8 GPUs	0.17

Inference is reported as seconds per forecast day per ensemble member. For WN2, training cost is reported for four independently trained models, each requiring 490 TPU-days. n.r. = not reported.

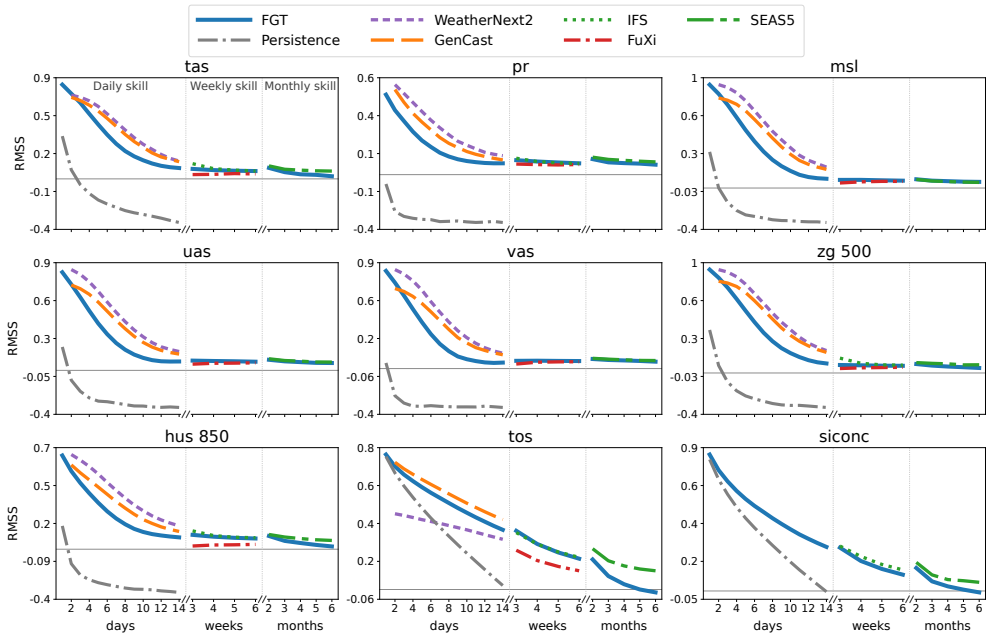


Fig. S3 Multihorizon deterministic forecast skill. The root-mean-square skill score (RMSS; higher is better) measures improvement over climatology across daily forecasts (days 1–14), weekly means (weeks 3–6), and monthly means (months 2–6) for nine atmospheric and surface variables. Daily comparisons pair raw, non-debiased FGT forecasts separately with WN2 ($n = 104$ matched initializations) and GenCast ($n = 101$); weekly comparisons pair debiased FGT with debiased FuXi-S2S ($n = 207$) and subseasonal IFS ($n = 301$); and monthly comparisons pair debiased FGT with the 51-member SEAS5 ensemble using $n = 78$ common initializations.

needed, at minimum, to reproduce the forecasts: initial conditions that preserve the model's native spatial interpolation scheme, coastal boundaries, and temporal aggregation for each variable; a model checkpoint; and inference code. We therefore compare FGT with SEAS5 at monthly lead times rather than combine non-equivalent forecasting experiments. The released FGT forecasts are intended to facilitate future AI seasonal benchmarks.

ClimateAR is a recent probabilistic model for seasonal-to-interannual forecasting [15]. The study compares performance relative to deterministic AI models and does not report CRPSS with respect to climatology. Nevertheless, Table S9 compares the empirical CRPS reported by ClimateAR with that of FGT for the four shared variables. Averaged over lead months 1–6, FGT outperforms ClimateAR by 39% for 2-m temperature, 21% for precipitation, 21% for mean sea-level pressure, and 21% for 500-hPa geopotential height. FGT's improvements over climatology are smaller than its improvements over ClimateAR. This is not a strictly like-for-like benchmark because the models were evaluated over different periods and with different ensemble sizes (ClimateAR has 200 members; FGT has 51). Notably, ClimateAR has approximately

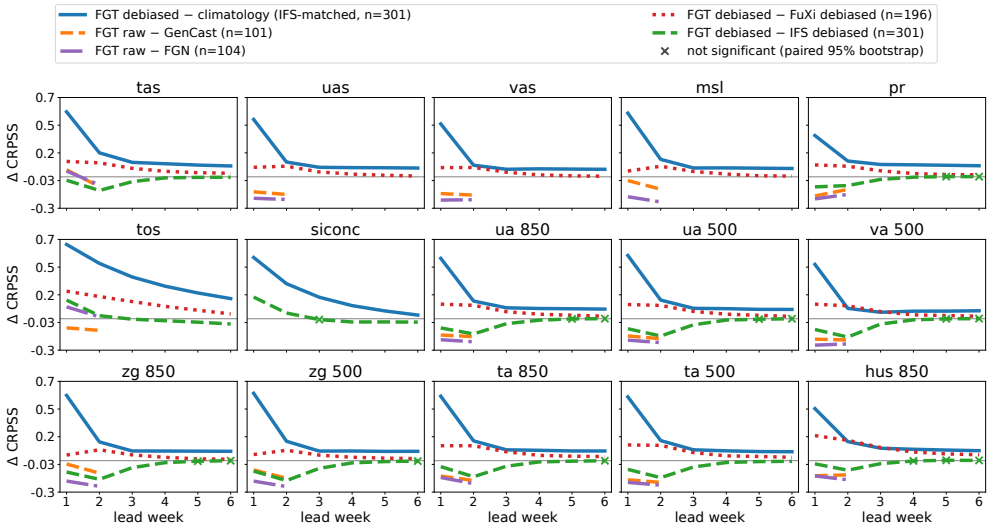


Fig. S4 Differences in probabilistic forecast skill matched over initialization dates, debiasing protocol, and approximately matched ensemble sizes. Weekly differences are defined as $\Delta\text{CRPSS} = \text{CRPSS}_{\text{FGT}} - \text{CRPSS}_{\text{reference}}$ and are shown across lead weeks 1–6 for 15 atmospheric and surface variables. Raw, non-debiased forecasts compare FGT with WN2 ($n = 104$ matched initializations) and GenCast ($n = 101$), while debiased forecasts compare FGT with FuXi-S2S ($n = 207$) and subseasonal IFS ($n = 301$). The debiased FGT skill relative to climatology is evaluated using the same $n = 301$ initializations as the IFS comparison. WN2 and GenCast are available through week 2; the subseasonal comparisons extend through week 6. Positive values favor FGT, and the horizontal zero line indicates equal skill. Crosses mark differences whose paired 95% bootstrap confidence interval includes zero, based on 1,000 initialization-date resamples.

480 million parameters and predicts monthly averages, whereas FGT has 35 million and resolves intervening daily averages.

4.3 Long-range prediction: stability versus skill

We evaluate a single trajectory highlighted by the TritonCast authors [23] (initialized on 1 January 2020). To make the comparison maximally favorable to TritonCast, monthly verification fields are reconstructed from its own released ERA5 trajectory by averaging the daily 12 UTC states [70]. All models are evaluated against this same reference, although the other systems produce monthly means using a different temporal-aggregation convention. Still, TritonCast has substantially higher RMSE than SEAS5, FGT, and the trend-aware climatology (Fig. S5). This shows that stability is an insufficient condition for producing skillful forecasts.

We additionally extend WeatherNext 2 Mini (WN2) [26], a state-of-the-art probabilistic medium-range weather model, to seasonal lead times. The released demonstration provides a single preprocessed HRES initialization compatible with the Mini model, on 7 October 2024. We restrict this experiment to that initialization also because generating a 51-member seasonal ensemble required approximately 10 hours on our hardware, compared with approximately 19 minutes for FGT. WN2 resolves

Table S8 Comparison with recent AI studies extending to seasonal lead times.

Model	Model output	Evaluation	Publicly available
Ref. [13]	3-month means for 5 atmospheric and surface variables	3-month means for 2 variables, 1 month after initialization	Code, but no forecast archive, model checkpoint, or prepared initial conditions
Ref. [19]	6-hourly atmospheric states with prescribed sea-surface temperature and sea ice	DJF means from ensembles initialized between late October and early November, for a restricted set of variables and indices	ACE code, model checkpoint, and DJF ensemble means for 3 variables, but no prepared initial conditions or seasonal pipeline
Ref. [20]	12-hourly atmospheric states with prescribed sea-surface temperature	DJF means from ensembles initialized on 1 November	GenCast code and weights, but no forecast archive, initial conditions, or seasonal pipeline
Ref. [21]	Coupled atmosphere–ocean trajectories	Qualitative analysis of long-rollout behavior and physical processes	No forecast archive, code, or checkpoint available
Ref. [23]	Deterministic year-long atmospheric trajectories	Selected long-range case studies, including the 1 January 2020 initialization	Code, checkpoint, and initial conditions
Ref. [15]	Monthly means	12 variables	Code and checkpoint, but no forecast archive or initial conditions
FGT	Free-running daily forecasts	56 variables over 6 monthly leads	Forecast archive released; code and checkpoint to be released

six-hourly rather than daily states, accounting for roughly a fourfold increase in autoregressive steps. Dividing the 10-hour runtime by this factor gives approximately 2.5 h; part of the remaining runtime difference relative to FGT could reflect running TPU-optimized inference code on a GPU. WN2 remains competitive with FGT and trend-aware climatology over the first few months, after which CRPS increases sharply for several variables (Fig. S6). Seasonal prediction lies well outside its intended forecasting task, but the experiment nevertheless shows that strong short-range forecasting performance does not automatically translate into useful long-term behavior.

Table S9 Empirical CRPS (lower is better) for variables reported by both FGT and ClimateAR. FGT values are averaged over monthly initializations from 2020–2026 with a 51-member ensemble; ClimateAR values are taken from its Appendix Table 5 and cover 2005–2014 for a 200-member ensemble.

Variable	Model	Lead month					
		1	2	3	4	5	6
tas (K)	FGT	**0.381**	**0.519**	**0.557**	**0.580**	**0.595**	**0.608**
	ClimateAR	0.831	0.847	0.878	0.900	0.920	0.939
pr (10^{-6} kg m^{-2} s^{-1})	FGT	**6.88**	**7.95**	**8.13**	**8.17**	**8.21**	**8.29**
	ClimateAR	10.1	9.96	10.0	10.1	10.1	10.1
msl (Pa)	FGT	**100**	**125**	**128**	**129**	**130**	**131**
	ClimateAR	155	155	158	158	159	160
zg500 (m)	FGT	**12.7**	**16.1**	**16.6**	**16.9**	**17.1**	**17.4**
	ClimateAR	19.8	19.9	20.4	20.6	21.0	21.3

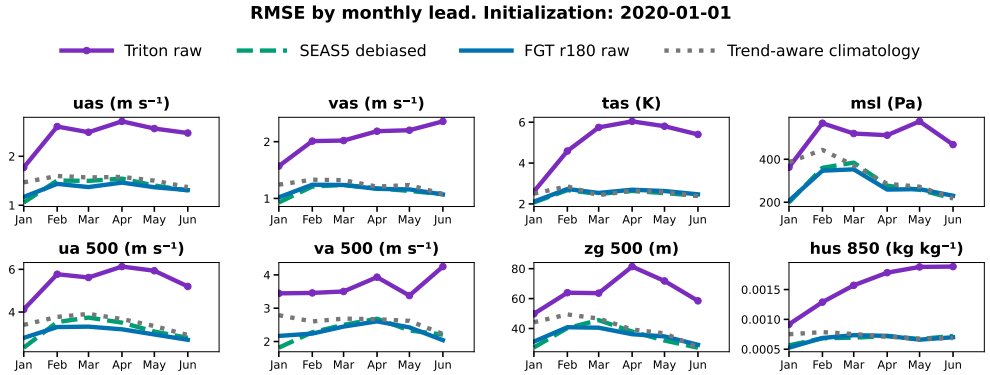


Fig. S5 Spatial RMSE by monthly lead for the showcased 1 January 2020 initialization. All models are evaluated against the same ERA5 reference used by TritonCast – monthly means are reconstructed from the released daily 12 UTC trajectory. Seasonal-model curves are computed using the ensemble mean. This experiment reproduces TritonCast's highlighted 2020 seasonal-forecasting case study.

5 Seasonal skill of FGT

5.1 ENSO precipitation teleconnections

To assess whether the predicted tropical-Pacific sea-surface temperature is translated into an atmospheric response, Fig. S7 shows precipitation anomalies regressed on the Niño 3.4 index. At lead month 2, FGT captures the tropical structure of the observed ENSO precipitation teleconnection. The predicted response is smoother than the ERA5 regression, however. At longer leads, an important distinction emerges between the two FGT regressions. When FGT precipitation is regressed on the ERA5 Niño 3.4 index, the teleconnection amplitude weakens substantially and is strongly

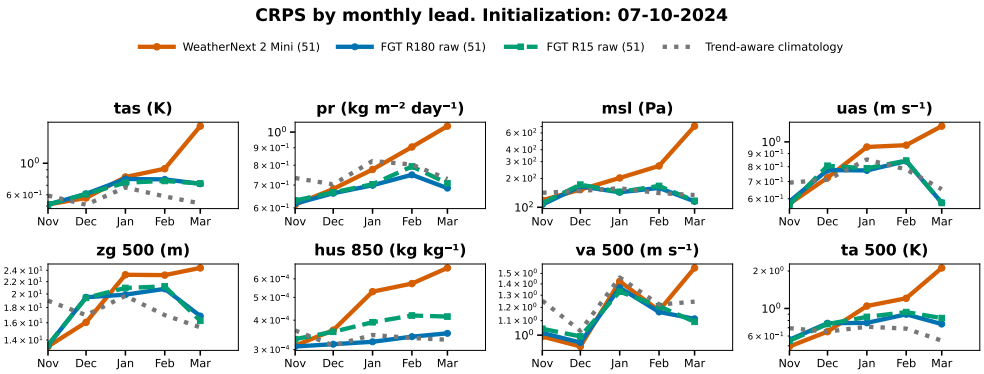


Fig. S6 CRPS by monthly lead for one initial condition, provided by WeatherNext2. WeatherNext 2 Mini, FGT R15, and FGT R180 use 51-member ensembles.

diminished by month 6. In contrast, regressing FGT precipitation on FGT's own predicted Niño 3.4 index produces a spatially coherent tropical response even at month 6. The failure stems from damping of the predicted Niño 3.4 anomaly, rather than from a collapse of the precipitation response. SEAS5 preserves both the Niño 3.4 anomaly and its associated precipitation pattern more effectively, resulting in a stronger forecasted teleconnection at month 6.

5.2 Spatial distribution of skill

We plot the spatial CRPSS and RMSS averaged over the test set in Fig. S8 and Fig. S9, respectively.

5.3 Global skill evaluation

The seasonal skill figures compare FGT against both SEAS5 and climatological reference forecasts across variables and lead months. Figs. S12–S16 summarize complementary aspects of forecast quality: deterministic error reduction, temporal and spatial anomaly correlation, probabilistic skill, and ensemble calibration.

These diagnostics examine three properties of seasonal performance: whether the ensemble improves on climatology, whether it captures anomaly structure after removing trends, and whether its spread is consistent with forecast errors.

5.4 FGT performance on additional unseen years

For each initialization month, RMSE and fair CRPS were computed against ERA5 for the 51-member FGT ensemble and then averaged over lead months 1–6 and the retained monthly variables. Metrics for the shaded historical period are obtained from a separately trained FGT for which the years 1983–1989 are held out from training, while the later segment corresponds to the network for which the years 2020–2026 are held out.

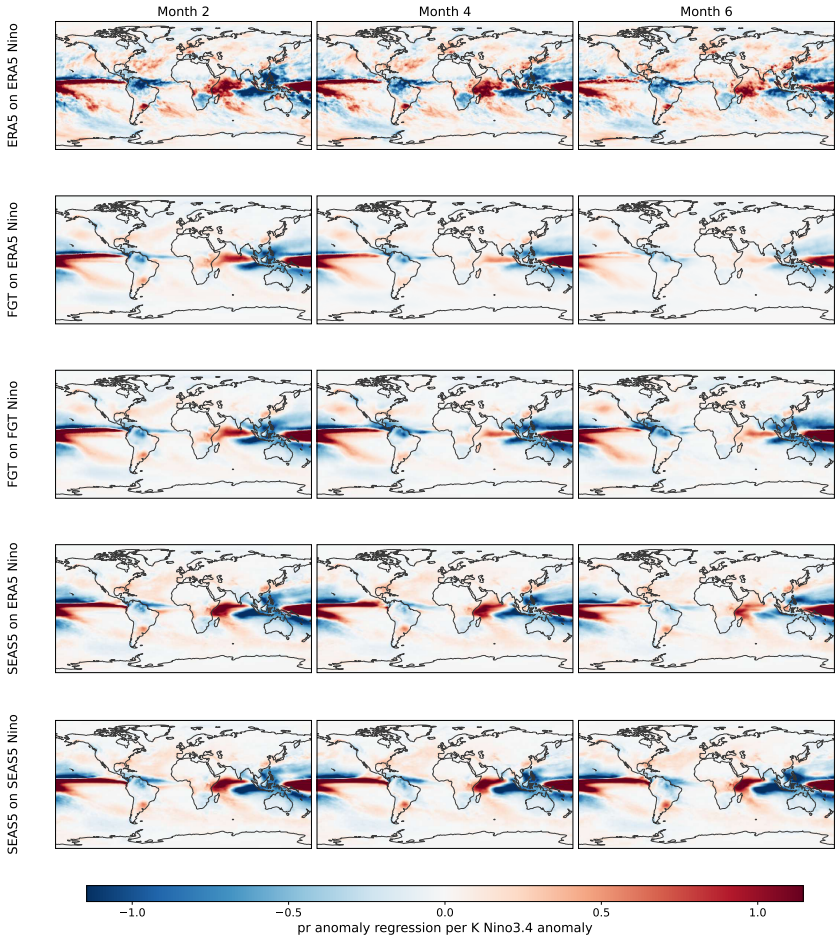


Fig. S7 ENSO teleconnection regression for precipitation anomalies, estimated jointly from all calendar months in the test set. Maps show the precipitation anomaly regression per 1 K Niño 3.4 anomaly at lead months 2, 4, and 6, comparing observed and forecast teleconnection structure across ERA5, FGT, and SEAS5.

6 Prediction snapshots

The prediction snapshots in Figs. S18–S25 provide qualitative examples of a 184-day FGT ensemble forecast initialized on 1 January 2020. These maps illustrate the physical structure and member-to-member diversity of the generated forecasts across atmospheric, oceanic, land, and sea-ice variables.

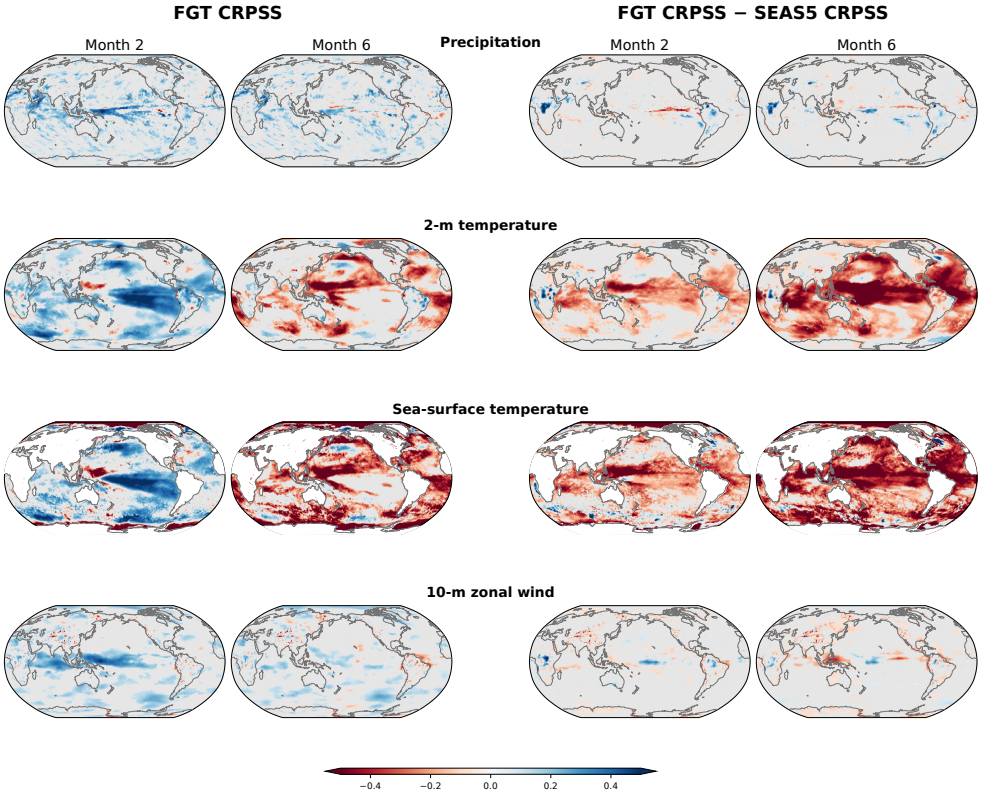


Fig. S8 Spatial probabilistic skill at lead months 2 and 6 for precipitation, 2-m temperature, sea-surface temperature, and 10-m zonal wind, averaged over the 2020–2026 test period. The two left columns show FGT CRPSS relative to the trend-aware climatological ensemble: positive values (blue) indicate lower CRPS than climatology, while negative values (red) indicate worse skill. The two right columns show the difference between FGT and SEAS5 CRPSS over matched initialization dates: positive values (blue) favor FGT, whereas negative values (red) favor SEAS5. Grey regions indicate grid cells for which the corresponding 95% bootstrap confidence interval includes zero and is therefore statistically insignificant.

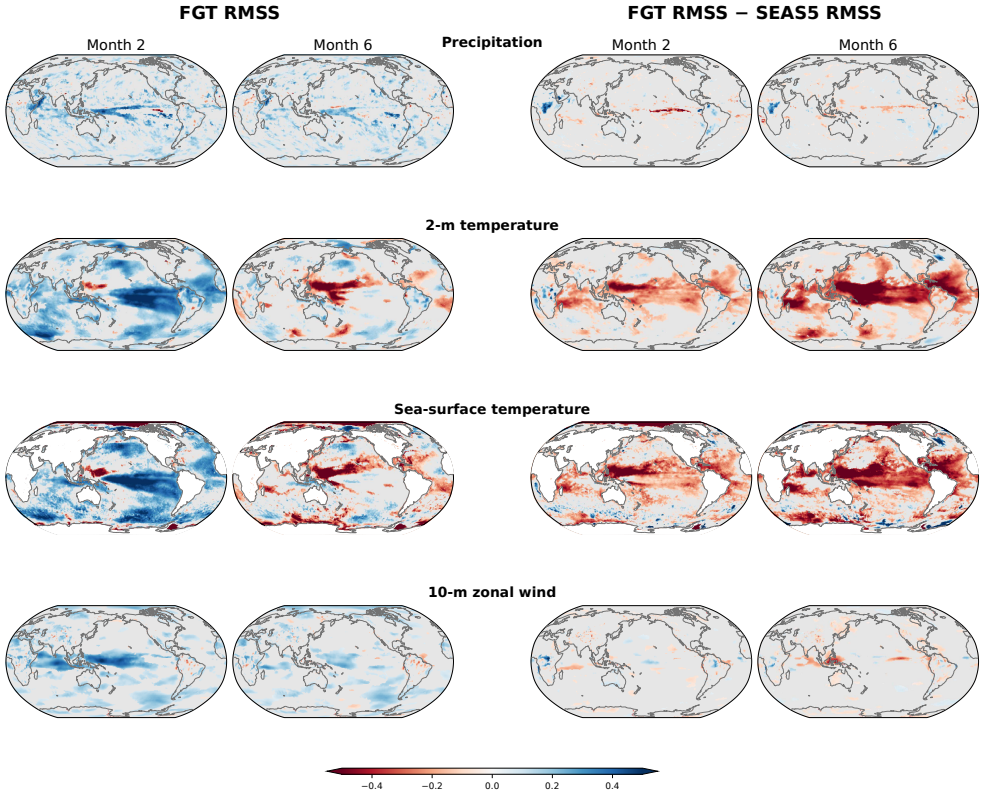


Fig. S9 Spatial deterministic skill at lead months 2 and 6 for precipitation, 2-m temperature, sea-surface temperature, and 10-m zonal wind, averaged over the 2020–2026 test period. The two left columns show FGT RMSS relative to the trend-aware climatological ensemble: positive values (blue) indicate lower RMSE than climatology, while negative values (red) indicate worse skill. The two right columns show the difference between FGT and SEAS5 RMSS over matched initialization dates: positive values (blue) favor FGT, whereas negative values (red) favor SEAS5. Grey regions indicate grid cells for which the corresponding 95% bootstrap confidence interval includes zero and is therefore statistically insignificant.

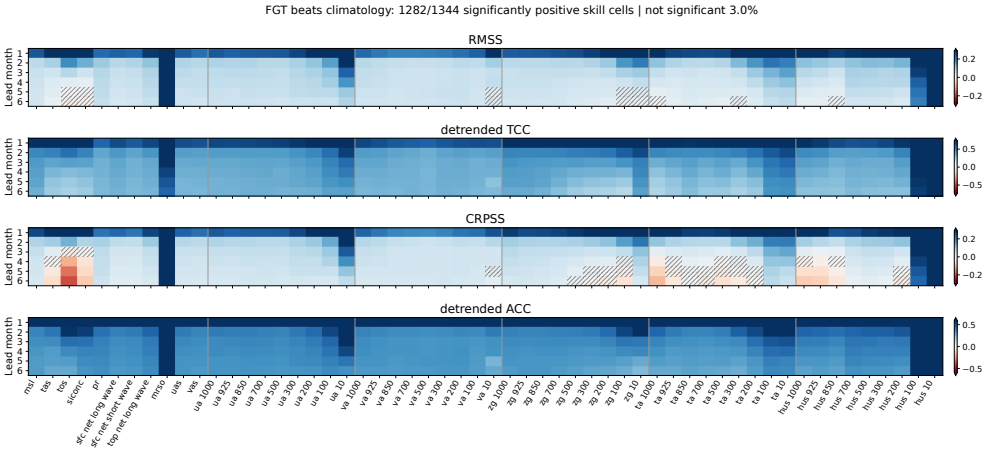


Fig. S10 Monthly skill of the 51-member FGT seasonal ensemble over the test-set period, shown by lead month and variable. Colors show mean skill relative to the climatological reference; blue indicates positive skill (greater than zero) and red indicates negative skill. Hatched cells are not statistically significant under a bootstrap over initialization dates, using a two-sided 95% interval against zero. We refer to Table S1 for the complete names of the variables.

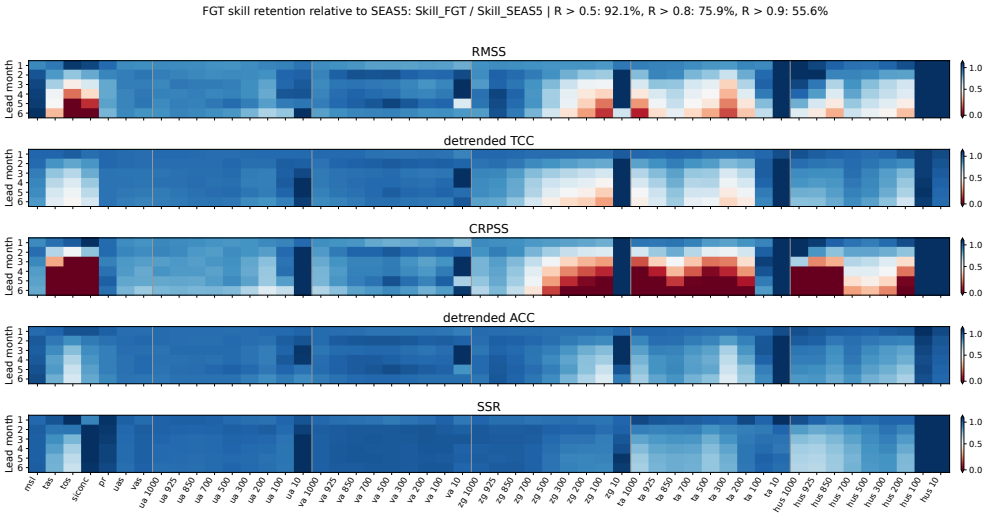


Fig. S11 Skill retained by the 51-member FGT seasonal ensemble relative to 51-member SEAS5 over matched initialization dates in the test set. Higher values favor FGT. For SSR, we first compute a calibration score indicating closeness to one and then compute the ratio of FGT to SEAS5 in the same way as other metrics. Consequently, a ratio higher than one would indicate that FGT has a better-calibrated spread than SEAS5. $R = 0$ denotes negative or close to 0 FGT skill while SEAS5 is better, and $R = 2$ denotes cases where SEAS5 has negative or close to 0 skill while FGT is better. We refer to Table S1 for the complete names of the variables. The colorbar is centered at 0.5, so that blue values indicate where FGT retains more than 50% of SEAS5 skill.

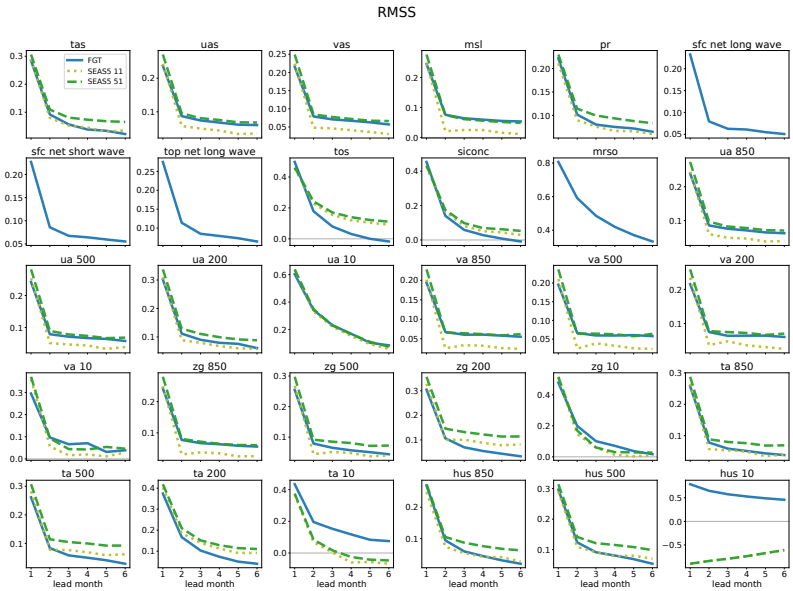


Fig. S12 Monthly root-mean-square skill score (RMSS) by lead month and variable. RMSS is computed on anomalies relative to the climatological reference, with positive values indicating lower error than climatology.

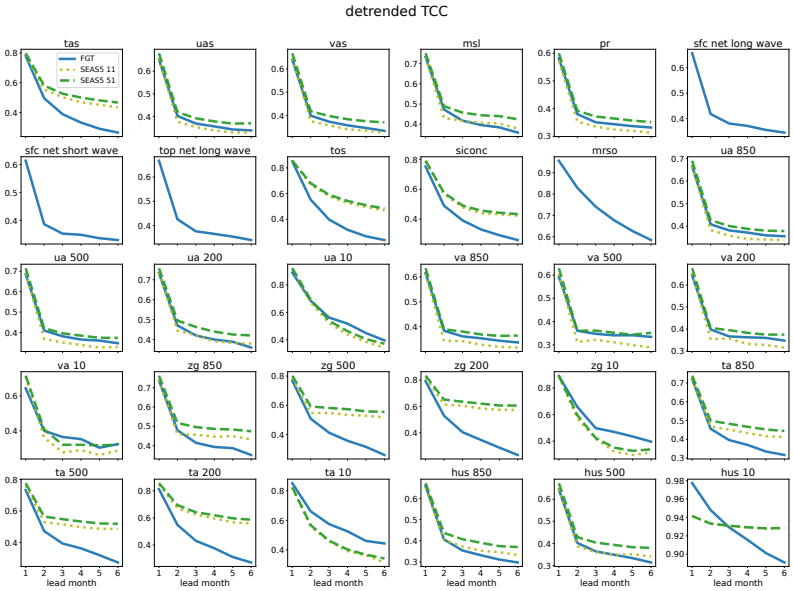


Fig. S13 Detrended monthly TCC of anomalies by lead month and variable. The trend is removed before verification, so the curves emphasize year-to-year forecast skill.

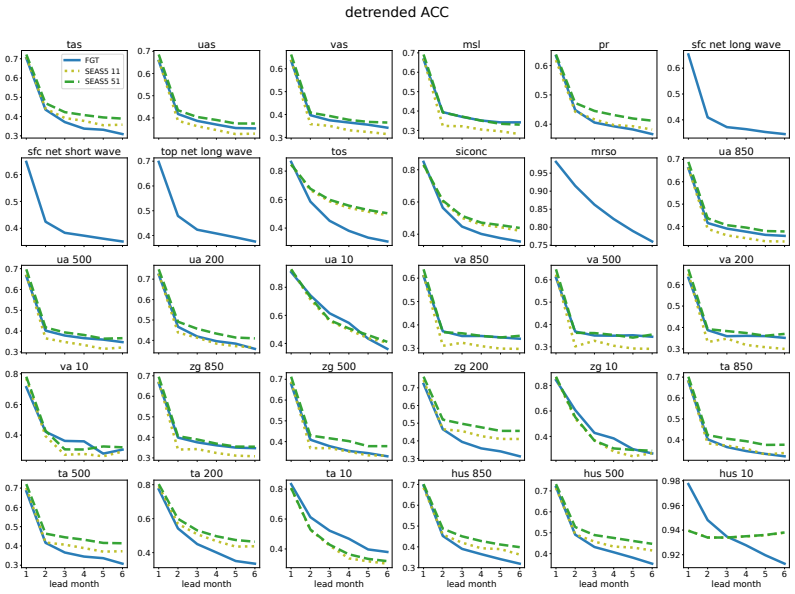


Fig. S14 Detrended monthly spatial ACC by lead month and variable. The trend is removed before verification, so the curves emphasize interannual anomaly skill.

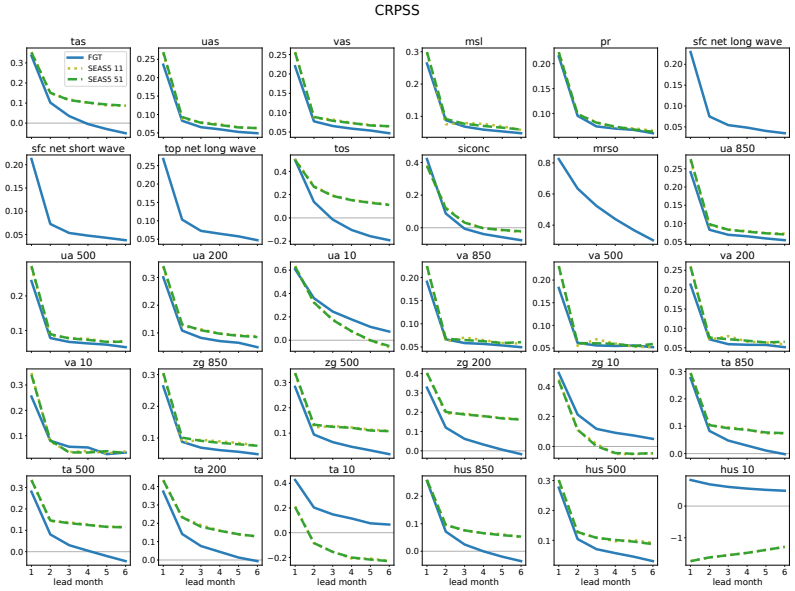


Fig. S15 Monthly seasonal continuous ranked probability skill score (CRPSS) on anomalies by lead month and variable. Positive CRPSS indicates probabilistic skill relative to climatology, while negative values indicate performance worse than the climatological ensemble reference.

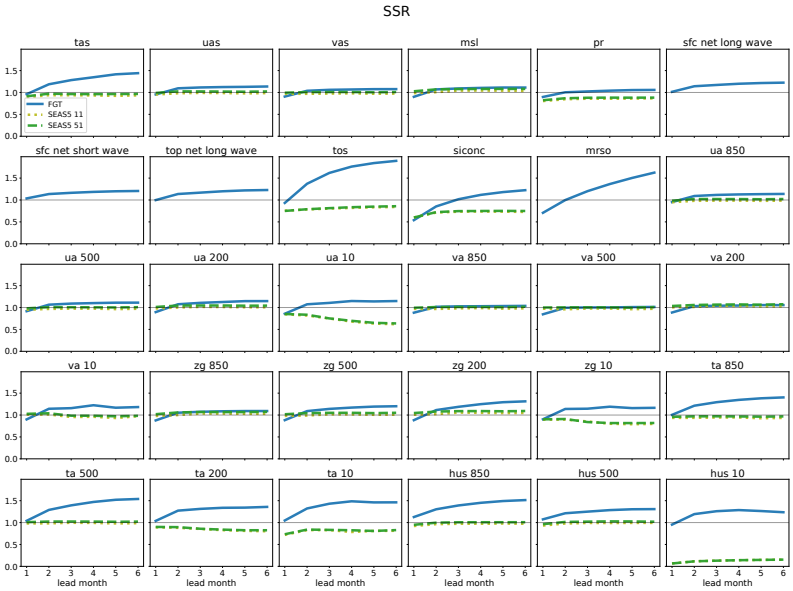


Fig. S16 Monthly seasonal spread-skill ratio (SSR) of anomalies by lead month and variable. Values near one indicate a calibrated ensemble spread relative to forecast error; values below one indicate underdispersion, and values above one indicate overdispersion.

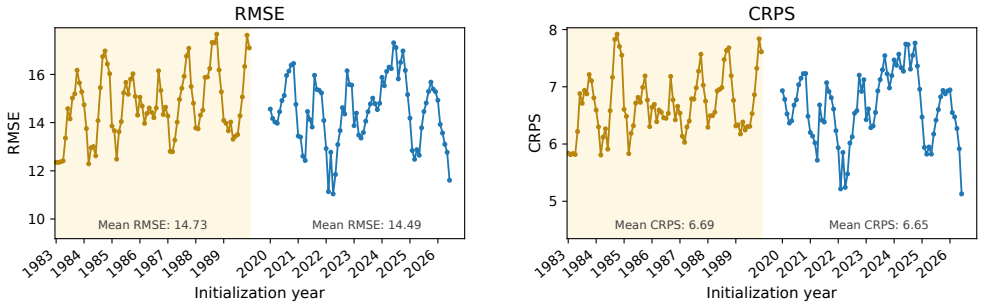


Fig. S17 FGT error by initialization month. Lines show the monthly sequence of FGT errors for the network trained with the years 1983–1989 held out and the network trained with the years 2020–2026 held out. Panels show root-mean-square error (RMSE) and fair continuous ranked probability score (CRPS). Values are averaged over lead months 1–6 and the retained monthly target variables on the 1° grid; lower values indicate better forecasts.

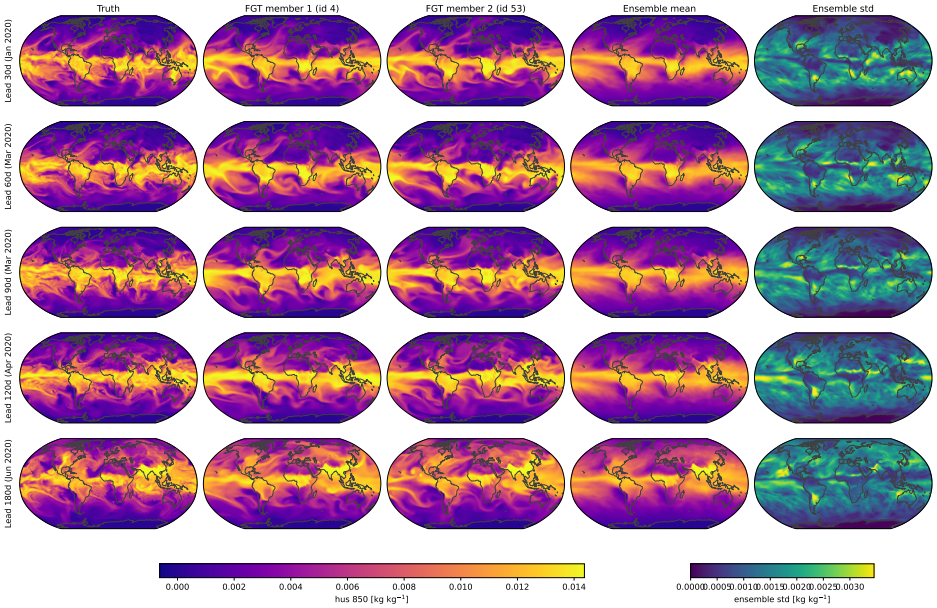


Fig. S18 Prediction snapshot for 850-hPa specific humidity from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

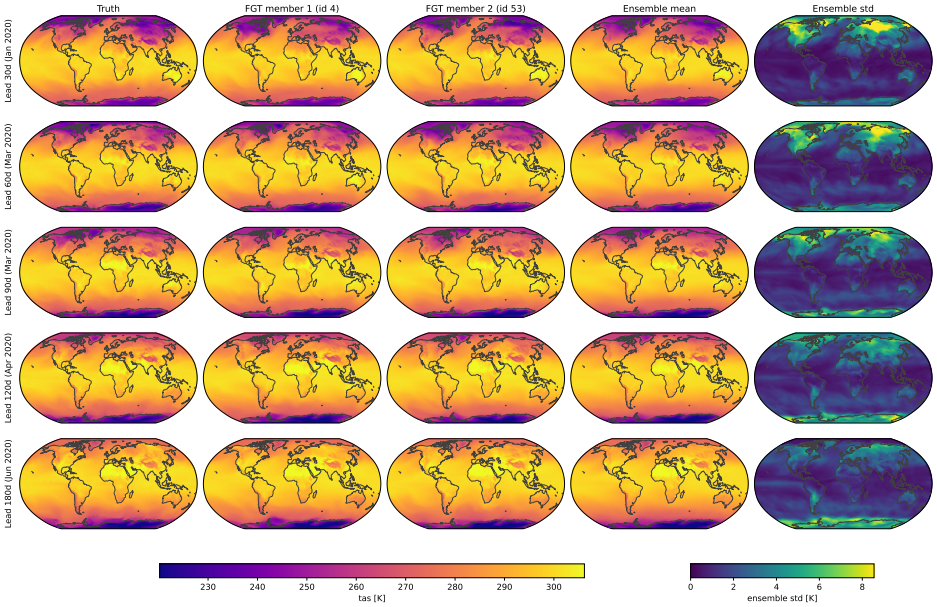


Fig. S19 Prediction snapshot for 2-m temperature from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

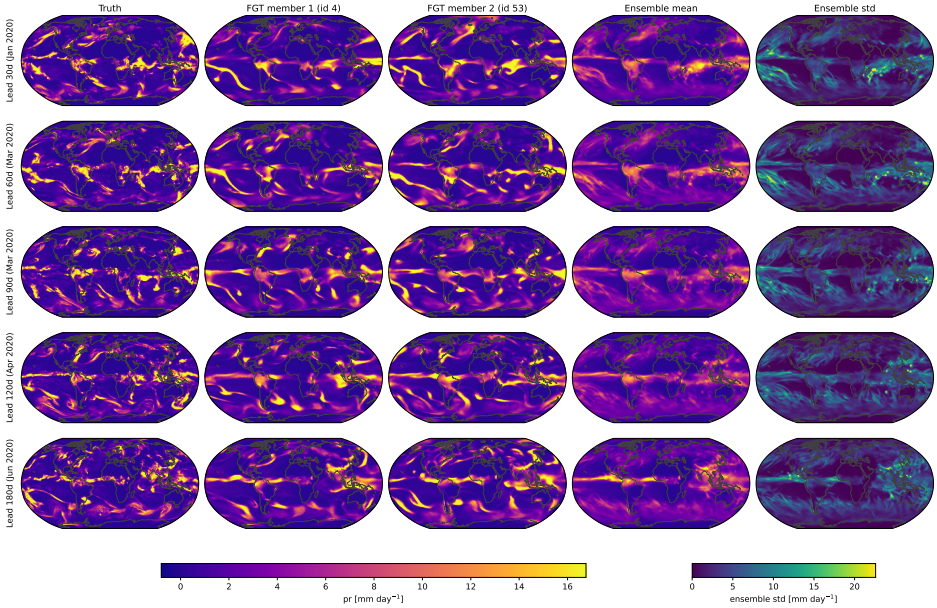


Fig. S20 Prediction snapshot for precipitation from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

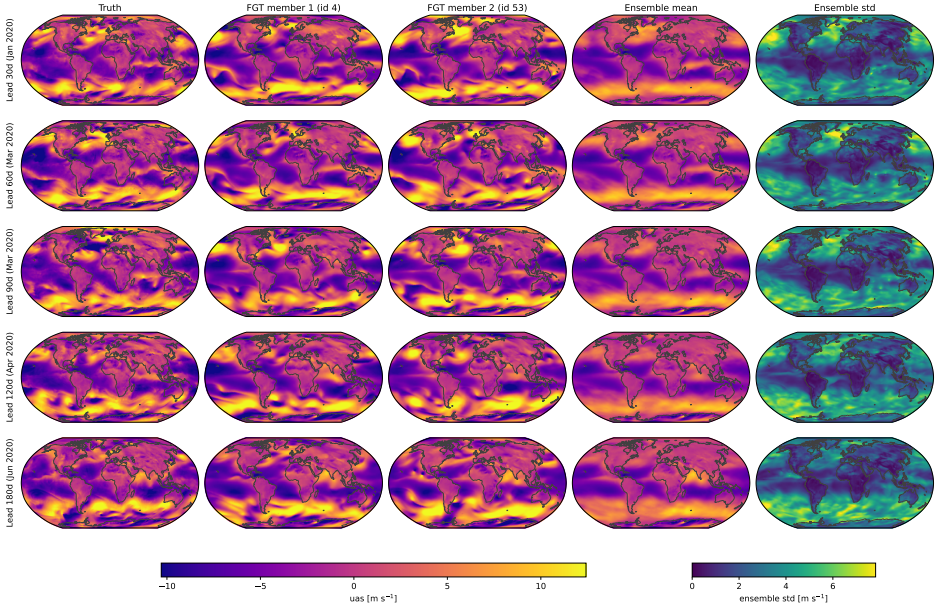


Fig. S21 Prediction snapshot for near-surface zonal wind from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

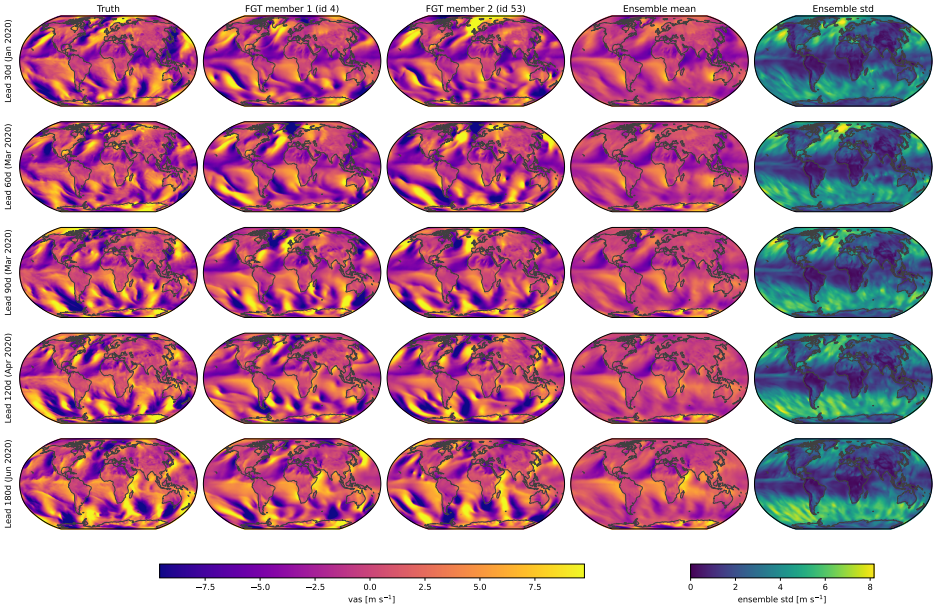


Fig. S22 Prediction snapshot for near-surface meridional wind from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

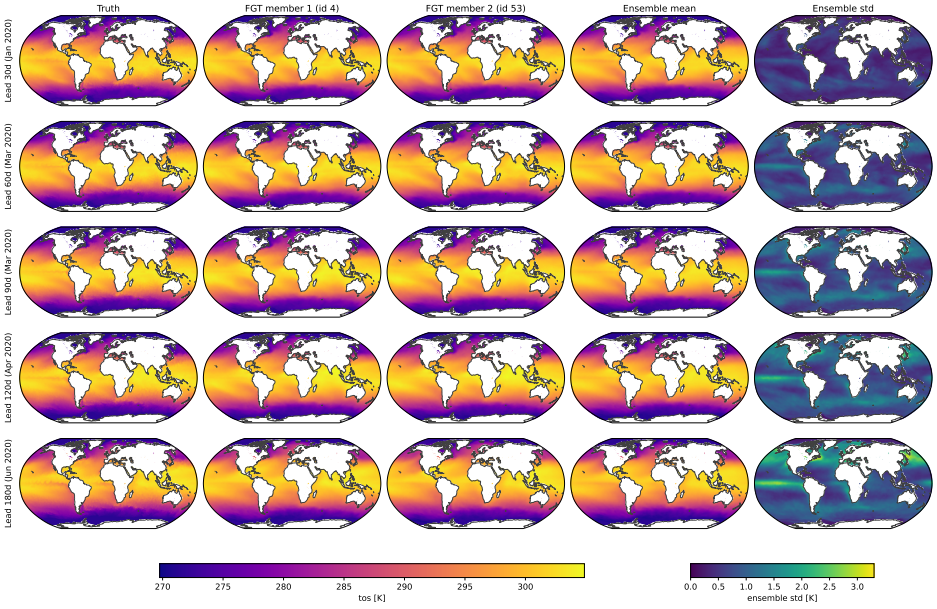


Fig. S23 Prediction snapshot for sea-surface temperature from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

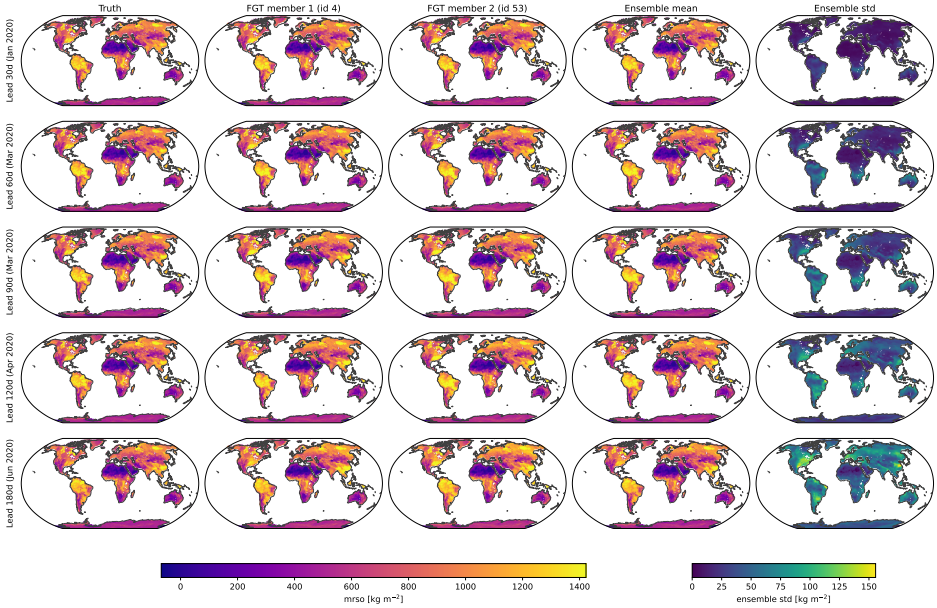


Fig. S24 Prediction snapshot for total soil moisture from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.

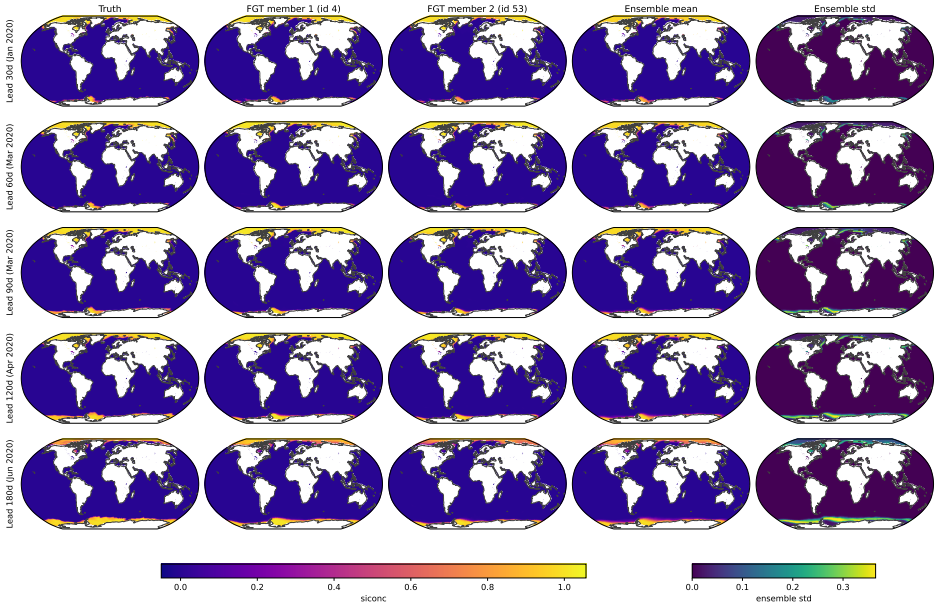


Fig. S25 Prediction snapshot for sea-ice concentration from a 184-day FGT ensemble forecast initialized on 1 January 2020. Panels show selected ranked ensemble members in physical units.