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.

Fine-tuning a weather model for improved ensemble subseasonal forecasting

Vateanui Sansine

vateanui.sansine@ird.fr

Santé et Services des Ecosytèmes Polynésiens

Takeshi Izumo

French National Research Institute for Sustainable Development

Marania Hopuare

Géopôle du Pacifique Sud

Damien Specq

Météo-France, CNRS, Univ. Toulouse, CNRM

Sophie Martinoni-La Pierre

Météo-France

Article

Keywords:

Posted Date: October 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10744980/v1

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

Additional Declarations: No competing interests reported.

Fine-tuning a weather model for improved ensemble subseasonal forecasting

Vateanui Sansine [1]*, Takeshi Izumo [2], Marania Hopuare [3], Damien Specq [4] and Sophie Martinoni-La Pierre [5]

1 UMR-241 SECOPOL, French Polynesia; vateanui.sansine@ird.fr

2 French National Research Institute for Sustainable Development, UMR-241 SECOPOL, French Polynesia; takeshi.izumo@ird.fr

3 GEPASUD, University of Polynesia, Campus d'Outumaoro, 98718 Puna'auia, Tahiti, French Polynesia; marania.hopuare@upf.pf

4 Météo-France, CNRS, Univ. Toulouse, CNRM, Toulouse, France; damien.specq@meteo.fr

5 Météo-France, Direction de la Climatologie et des Services Climatiques, Toulouse, France; sophie.martinoni@meteo.fr

* Correspondence: vateanui.sansine@ird.fr

Abstract:

Subseasonal forecasting (15 to 42 days ahead) is socio-economically critical yet remains challenging. Recently, machine-learning based global weather forecasting could show superior skill than traditional forecasting systems. Yet, training these data-driven AI models require immense computational resources. Here we demonstrate, by using a medium-range model, GraphCast, as pre-trained model and focusing on reducing its iterative error accumulation, that fine-tuning is a cost-effective strategy to improve subseasonal forecasting. We further introduce a simple and computationally-inexpensive ensemble method: creating 11 variants of our fine-tuned model, FT, through different data sampling during fine-tuning and considering each as a member of a new model, FT ensemble. This model significantly surpasses the ECMWF Subseasonal to Seasonal ensemble model (S2S ensemble) for week 3 and week 4 lead-times, notably for precipitation, crucial for water management. These results highlight that combining the proposed fine-tuning and ensemble methods are powerful, low-cost pathways to improve atmospheric and ocean forecasting.

Introduction

Subseasonal forecasting is the process of predicting weather patterns that occur between 2 to 6 weeks or more in the future[1]. It fills the gap between medium-range forecasting, that extends up to 15 days, and seasonal forecasting that gives an outlook of the coming trends months ahead. A skillful subseasonal forecast is critical across multiple society relevant sectors[2] and for pro-active disaster impact mitigation efforts, since these efforts may take several weeks to implement[3].

The difficulty of implementing accurate subseasonal forecasts originates from the dampening and the loss of atmospheric initial conditions beyond a sufficiently long lead time[4], iterative error accumulation, as well as the slow changes in boundary conditions such as Sea Surface Temperatures (SSTs), soil moisture and sea-ice components. It is those different time and space scales of atmosphere, land and ocean, and the ability to predict them, that makes subseasonal forecasting a major challenge[5,6,7]. This timescale has long been seen as a "predictability desert"[8] by meteorologists. However, there are important potential sources of predictability for subseasonal timescales to be found in certain large-scale phenomena such as the intraseasonal Madden-Julian Oscillation (MJO) and the interannual El Niño Southern Oscillation (ENSO).

Recent advancements in machine learning models represent a complementary alternative to the "traditional" dynamical models which are based on fluid mechanics and thermodynamics equations, for weather forecasting at medium-range[9,10,11,12], subseasonal-range[13,14] , and for long-term weather and climate forecasting[15]. The data-driven medium-range weather forecasting models only utilize machine-learning techniques and were trained using about 40 years of historical data obtained from the European Center for Medium-Range Weather Forecasts (ECMWF) reanalysis v5 (ERA5)[16]. They achieved notable success in forecasting skillfully weather variables globally for up to 10-days ahead, while requiring only a fraction of the computational resources used by traditional dynamical models[9,11].

At longer lead-times, i.e. for subseasonal forecasts, machine learning models have also made significant improvements[17,18,19]. However, there are some challenges and limitations[20,21], notably arising from the incorporation of a limited amount of variables into the models and, possibly more importantly, from the iterative error accumulation.

One of the interesting novelty of the recent medium-range models[9,11,12] is that they have a comprehensive amount of meteorological variables compared to earlier studies with a limited number of variables. Indeed, FourCastNet[12] has 20 upper-air atmospheric variables at 5 pressure levels, Pangu-Weather[11] has 5 variables at 13 pressure levels plus 4 surface variables and Graphcast[9] possesses 6 variables at 13 pressure levels plus 5 surface variables. However, the evaluation of such models has been limited to a 10-day lead-time. Yet, we expect their forecasting skills to rapidly decrease with lead-time, due to iterative error accumulation. Our first objective was, after assessing their accuracy for subseasonal forecasts, to

identify methods to enhance their precision at week 3 and week 4 lead-times. We have thus developed an efficient fine-tuning process, that leads to improvements in forecasting skills, even at the 4-week lead, while remaining computationally cost-effective.

Fine-tuning in artificial intelligence (AI) is the procedure of adjusting a pre-trained model and adapt it to a new and specific task. When a pre-trained model is fine-tuned, it uses its knowledge acquired from the original task to learn the nuances of the new task more efficiently. The primary hypothesis of our work is that, by fine-tuning a medium-range model to predict directly the weekly means, the long-term prediction errors can be decreased. Consequently, the obtained results can surpass those of their non-fine-tuned counterparts, with a much higher computing efficiency than if the model was developed and trained from the very beginning.

The process of fine-tuning machine learning models for weather forecasting has already been applied to medium-range timescales, for example 3 pre-trained FuXi[22] models were fine-tuned for optimal forecast performance for one of the forecast time windows: 0-5 days, 5-10 days, and 10-15 days. Additionally, the Graphcast operational model has undergone fine-tuning using HRES data from 2016 to 2021. Nevertheless, fine-tuning has not yet been accomplished for subseasonal timescales. In the present study, we utilize Graphcast as a pre-trained global weather model ("REF") and adjust it through fine-tuning to get significantly-improved weekly-mean global subseasonal forecasts at a 100 km resolution (1°).

The benefit of this strategy is the ability to generate precise deterministic subseasonal forecasts with minimal computational power required for training and without the need for creating a new model structure, which is a time-consuming procedure. This strategy exploits the existing knowledge of the medium-range pre-trained model, adjusting it to longer timescales.

The proposed fine-tuning strategy adjusts Graphcast by substituting its standard input and target data (see Methods and related Fig. 4). As input data, Graphcast utilizes the two of the most recent Earth's atmospheric states (from ERA5), i.e the current time (t0) and 6 hours earlier, and forecasts the next atmospheric state 6 hours ahead. We substituted the '6 hours earlier' input with the 'preceding weekly mean', i.e. the mean value from the preceding 7 days. Hence, the input data consists of the ERA5 mean of the preceding week and the 6-hourly one at time t0 (the initial time from which the forecast starts). Concerning the target data, we substituted it with the weekly mean value of the week to be predicted. This allows the model to retrain on the relevant timescale of interest for subseasonal forecasts with the advantage of removing iterative error accumulation originating from usual roll-out forecasts (roll-out forecasts being produced following the principle of feeding the forecast back to the model to have a longer lead-time forecast[9]). One of the main advantage of predicting directly week 2, 3 or 4 is that there is less error accumulation than on a standard roll-out forecast. Indeed, in order to obtain subseasonal forecasts from the normal version of Graphcast, we would have needed to perform for example 56 roll-out forecast (4 per day)

to have an idea of the weather 14 days ahead. With FT, one forecast is needed to obtain a 14-day ahead forecast for example.

Here we first implement data-driven subseasonal forecasts that predict weekly-mean weather conditions for the second (7-14 days ahead), third (14-21 days ahead) and fourth (21-28 days ahead) week in advance, here referred to as week 2, week 3 and week 4, using a fine-tuned version of Graphcast, named "FT".

Then we implement ensemble forecasting, by creating 11 different versions of FT, each trained on a different subset of the 4-year ERA5 dataset (Table 1), and by considering each FT variant as a member of a new ensemble model, named "FT ensemble". This allowed us to enhance the accuracy of our forecasts using a simple, efficient and cost-effective computational method (see Methods).

Our ensemble fine-tuned model, FT ensemble, was compared to the reforecasts of the ECMWF Subseasonal to Seasonal (S2S) system, which is recognized as the best ocean-atmosphere modeling system for producing subseasonal forecasts[23,24]. Despite certain limits and areas for improvement, we argue that combining the proposed fine-tuning and ensemble methods to medium-range weather models is a viable and efficient strategy for significantly increasing accuracy in subseasonal forecasts with low computational and human-resources costs.

Results

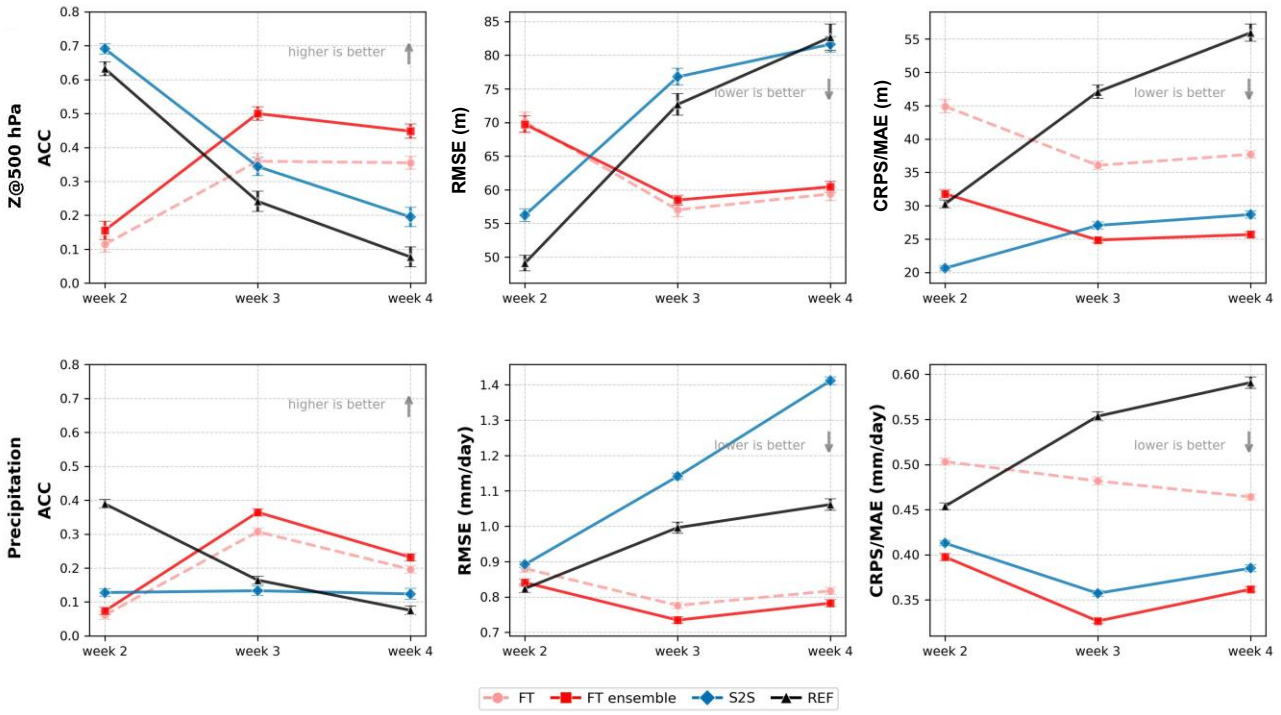
Here we first compare the performances of subseasonal forecasts from our FT ensemble and 'deterministic FT' (i.e. only the control member fine-tuned on the longest training period, 2016-19; cf. Methods) models with our reference, "REF" (Graphcast Operational), which is also deterministic, and reforecasts (i.e. hindcasts) from S2S ensemble. The forecasting models are evaluated throughout the 3-year 2020-2022 period (which was not used for training/fine-tuning), with three forecast lead-times: week 2, 3 and 4. For each lead-time, the daily forecasts are reduced to their weekly mean before being evaluated against observations, except for the FT models which were fine-tuned to produce weekly mean forecasts directly.

Figure 1 displays the Anomaly Correlation Coefficient (ACC), i.e the temporal mean of the spatial correlation between forecasts and observations, the Root Mean Squared Error (RMSE) and the Continuous Ranked Probability Score (CRPS; a metric relevant for ensembles; for deterministic models, CRPS is mathematically the Mean Absolute Error; MAE) for FT ensemble, deterministic FT, REF and S2S ensemble for all lead-times for the geopotential height at 500 hPa (Z500) and precipitation. See Supplementary Figure S1 and S2 for variables

such as 2-m air temperature (T2M), zonal wind U at 10 m, 500 and 850 hPa, Z200, Z850 and specific humidity at 200 and 850 hPa.

For week 2 lead-time, the best model is S2S ensemble in terms of ACC, RMSE and CRPS, except for the ACC for the precipitations where REF shows slightly better results. This underperformance from FT ensemble and deterministic FT could be explained by two facts. First, the standard training of REF is already sufficient and well suited for week 2 (which still stands in the medium-range lead-time, and for which the iterative error accumulation is reasonable). Second, our fine-tuning strategy for FT, by using as input the previous week's mean in addition to t0 initial state is more suited for longer lead-times. This weekly mean might be a too strong temporal averaging, losing short timescale information, to be relevant for week 2 lead.

However, for weeks 3 and 4, our fine-tuning strategy significantly improves REF. FT ensemble is the most accurate model, even better than S2S ensemble, for all metrics for both Z500 and precipitation (and for other variables; cf. Figs S1 and S2). In terms of ACC and RMSE, it is followed by deterministic FT, S2S ensemble and REF. In terms of CRPS, FT ensemble remains the best and is followed closely by the S2S ensemble. The improvement is the largest at week 4, with notably an about 2.3-fold increase in ACC (i.e. about 5-fold increase in explained variance) instead of about 1.5-fold at week 3, for Z500 from S2S to FT ensemble. Proving that our fine-tuning strategy, when combined with our ensemble modeling approach, can significantly enhance subseasonal predictions of medium-range weather models like our REF (Graphcast). These results also confirm the hypothesis that the prediction errors in models such as Graphcast strongly increase with forecast lead-time because of the roll-out strategy, and that they can be significantly reduced via our fine-tuning process and further via our ensemble forecasting method for longer lead-time forecasting. However, in terms of RMSE, the ensemble strategy has more ambiguous results (e.g. in Figs S1 and S2, we can see that FT outperforms FT ensemble for U-winds at 500 and 850 hPa at week 4, though not in terms of ACC and CRPS). This suggests that the ensemble method does improve the correlation (through ACC) and CRPS but the ensemble mean of FT ensemble might still be quite biased compared to FT. Still, we can conclude that, overall, our proposed fine-tuning method is really appropriate for extended lead-times, i.e. weeks 3 and 4.



187

Figure 1. ACC (left), RMSE (middle) and CRPS (right; equivalent to MAE for deterministic models with only one member, cf. Methods) forecasting skill metrics for REF (black lines and triangles) and FT (pink dashes and circles) deterministic model, and for FT (red lines and squares) and S2S (blue lines and diamonds) ensemble models (11 members each), for all lead-times forZ500 (m, upper) and precipitation (mm.day^{-1}, lower). See Supplementary Figure S1 and S2 for other variables.

188 The previous metrics were global. To have a better idea of the models' performance on a
189 local scale, we first show, as an example, T2M in Paris, for week 3 and week 4 lead-times
190 (see Supplementary Figure S3). This suggests how the fine-tuning strategy coupled with the
191 ensemble forecasting method can improve T2M forecasts on a local scale.
192
193 The natural question that arises is: how does the spatial distribution of FT ensemble skill
194 compare to that of REF, deterministic FT and S2S ensemble? Figure 2 shows, at each spatial
195 grid point of earth, the temporal correlation coefficient (R_t) for Z500 for all the
196 aforementioned models, as well as the difference in R_t between deterministic FT and REF,
197 and the difference between FT and S2S ensemble means for week 4 lead-time (see
198 Supplementary Figures S4 to S9 for Z500 at week 3 and for other variables such as T2M,
199 precipitation, zonal wind U at 10m, 200 and 850 hPa for week 3 and 4 lead-times separately).
200
201 First, Figure 2 upper panels show that, while REF performs poorly with a global mean $R_t =$
202 0.05, deterministic FT performs very well with a mean $R_t = 0.40$ for Z500 for week 4.
203 Demonstrating how much our fine-tuning method improves the accuracy of REF over land
204 and ocean in nearly all regions (FT minus REF: $\Delta R_t = 0.35$ for the global average). Notably
205 the equatorial Pacific experiences a large ΔR_t improvement of about 0.6-0.8 (a few regions
206 have not undergone any improvements, likely because of noise associated to the relatively
207 low signal/noise ratio for such point-wise metric).
208

Secondly, Figure 2 lower panels show that FT ensemble ($R_t = 0.46$) performs even better than FT in nearly all regions, and better than S2S ensemble in general ($R_t = 0.24$). However, for Z500, even though FT ensemble demonstrates superior skill than S2S ensemble from the subtropics to the poles, its performance in the tropics remains behind the S2S ensemble. This is due to the high correlation between Z500 and SSTs, which is mediated by the temperature of the layer between the surface and 500 hPa in the tropics[25]. The S2S ensemble model excels in capturing this relationship because it incorporates full ocean physics. Since tropical deep convection is tightly coupled to the warmest SSTs, this strong ocean-atmosphere coupling allows S2S to more skillfully forecast the diabatic heating that drives tropical temperature anomalies and, consequently, Z500 patterns. T2M is also strongly correlated with the SSTs in the tropics where S2S ensemble also outperforms FT ensemble for this particular variable over tropical oceans, but not over tropical lands where FT ensemble is overall still better (See Supplementary Figure S4). In contrast, FT lacks an active ocean model. It must infer the oceanic influence indirectly through lower atmospheric variables like T2M, which, while correlated with SST, cannot fully represent the dynamical ocean processes that govern SST evolution and its coupling to convection. Therefore, in the tropics—where Z500 is most directly a passive tracer of tropospheric temperature forced by coupled ocean dynamics—the S2S framework presents a more challenging benchmark to outperform for such variable. Still, FT ensemble remains better than S2S in the tropics for other socio-economically relevant variables such as precipitation (Fig. S5), crucial for water management, and lower troposphere specific humidity (Fig. S9), crucial for thermal comfort and heat risk forecasting (with metrics such as heat index or wet-bulb temperature).

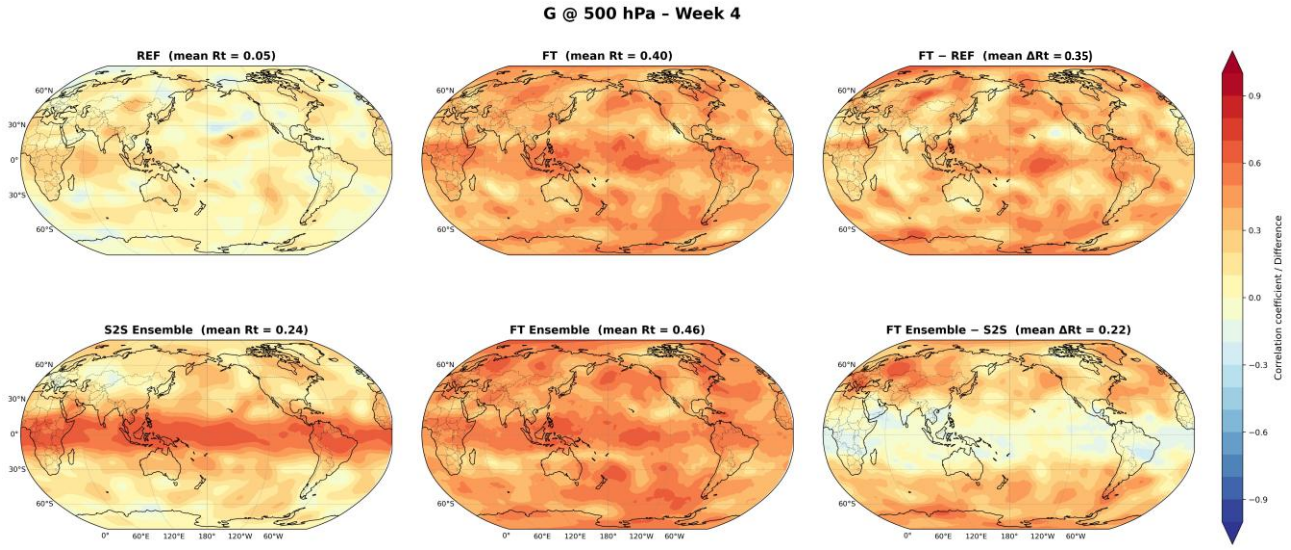


Figure 2. Comparison of the forecasting skill measured by the temporal correlation coefficient R_t, in upper panels between REF and FT deterministic models, and in lower panels between S2S ensemble and FT ensemble models, with the right panels showing their respective differences (the redder the better FT skill is), for Z500 for week 4 (21 to 28-day lead-time; for FT and S2S ensembles, ensemble means were calculated before computing R_t). The (latitude-weighted) global mean of R_t is indicated above each panel. See Supplementary Figure S4 for week 3 lead-time, and Figs S5 to S9 for other variables such as precipitation.

To synthetize the evolution of the skill as a function of latitude, Figure 3 illustrates the latitudinal distribution of the skill, measured by $R_{x,t}$ (cf. Methods), for the FT ensemble, S2S ensemble, and REF ensembles at week 4, for several variables. The $R_{x,t}$ assesses the skill of a model to predict together temporal and longitudinal variability. It is evident that the FT ensemble outperforms all other models for all variables, with the exception of Z500 and T2M near the equator where S2S ensemble is superior, as previously explained. The forecasting skill enhancement in the tropics as compared to extratropics, related to the strong atmosphere-ocean coupling, is more pronounced for S2S than for the atmosphere-only REF and FT models in all variables (except for precipitation), suggesting a large potential of improvement in future versions of REF and FT coupled to ocean. Still, overall, FT ensemble is the most skillful model, the second-best being the S2S ensemble, followed by REF.

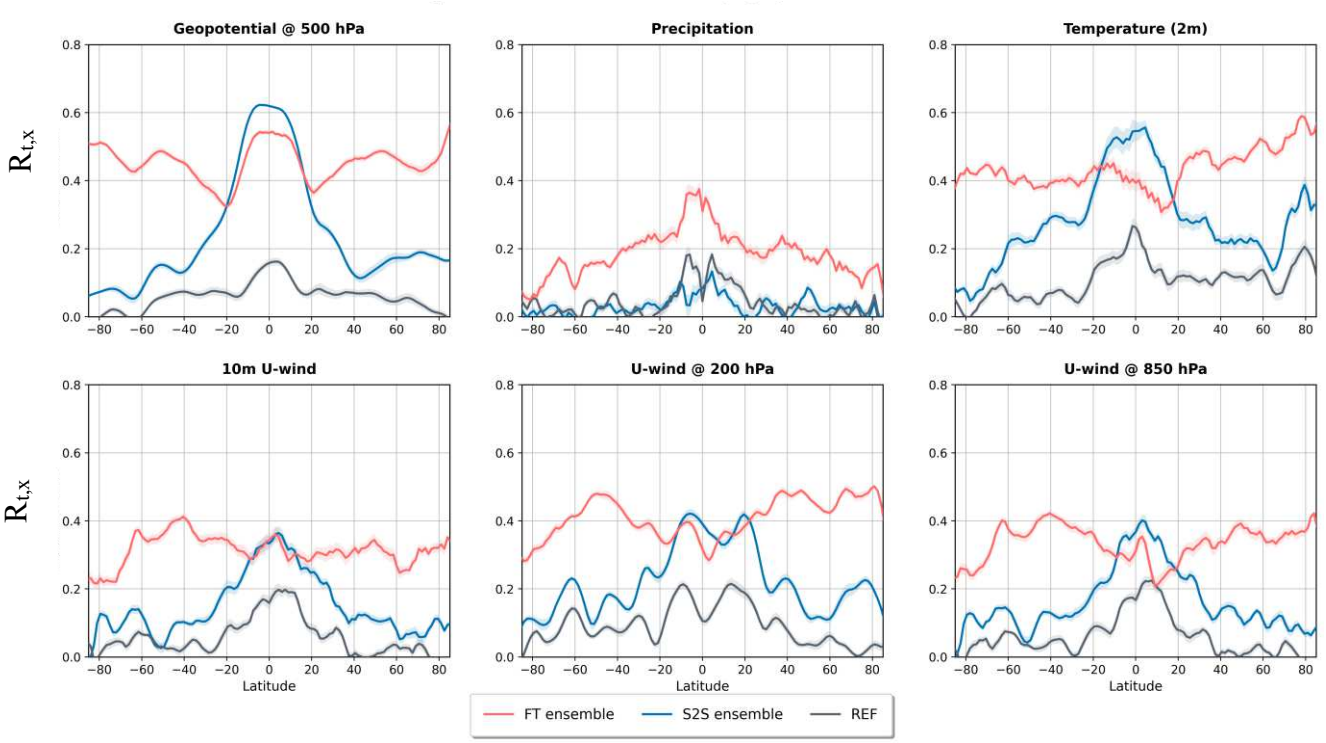


Figure 3. Latitudinal distribution of the forecasting skill, as measured by $R_{t,x}$ (cf. Methods), at week 4 for FT ensemble (red line), S2S ensemble (blue) and REF (black) for Z500, precipitation, surface air temperature at 2m, zonal wind U at 10m, 200 hPa and 850 hPa. See Supplementary Figure S10 for week 3 lead-time.

Discussion

Our proposed fine-tuning strategy coupled with our ensemble forecasting technique applied to a medium-range global weather forecasting model (Operational Graphcast = REF), has significantly increased its global performance on subseasonal timescales for all considered variables for week 3 and week 4 lead-times. Indeed, for these lead-times, FT ensemble performs better than deterministic FT, itself better than its "not-fine-tuned father" REF, and even better than the S2S ensemble model (the best traditional model at present for

intraseasonal forecasts) globally (except near the equator for geopotential height and surface temperature over tropical oceans where the S2S model is more accurate due to its ocean-atmosphere component).

An inherent benefit of the described fine-tuning plus ensemble procedure is its computational cost-effectiveness. Indeed, the fine-tuning process was carried out over a much shorter training period of 4 years (2016-2017-2018-2019), requiring much less data than the usual multidecadal training periods, and using only a 40 Gb A100 GPU. The training process took only around 24 hours per version of FT, meaning a total of 11 days of fine-tuning, and yielded promising results, overall outperforming S2S ensemble. A more comprehensive comparison can be drawn with REF (GraphCast at 0.25°), a model that underwent training on a 32 Gb Cloud Tensor Processing Units (TPU) v4 device, which typically performs about 15 to 30 times better than a GPU, for a duration of around 4 weeks[9]. REF employed a parallelization of data distribution approach that involved 32 TPU devices. Furthermore, they also employed Gradient check-pointing[26] to further decrease the memory usage.

The efficiency of our fine-tuning approach has enabled us to train 11 different versions of FT to provide ensemble forecasts at least similar to the ECMWF S2S ensemble system which contains 11 members. FT's cost-effectiveness may allow us to have much more members than the 11 members and further increase the accuracy of our ensemble forecasts (that already might be useful to capture weather extremes such as the record-breaking Siberian heatwave in 2020; Fig. S12). Though they use different techniques for ensemble forecasting, Pangu-Weather[11] generates a 100-member by applying Perlin noise to the initial states, and FourCastNet[12] uses a Gaussian random noise and added it to the initial conditions using Ensemble Kalman Filtering[27] to obtain a 100-member ensemble forecast. The ECMWF also introduced their ensemble Artificial Intelligence Forecasting System (AIFS)[28] based on optimizing a probabilistic proper score objective, which helps the model to represent model uncertainty, through shaping Gaussian noise.

It would also be interesting to extend the lead-time of FT ensemble to week 5 and week 6 as changing the target of the model and adapting the fine-tuning process is fairly easy and does not involve any architecture modifications. Indeed, FuXi-S2S[13], a state-of-the-art machine learning subseasonal range weather forecasting model, provides skillfull data-driven subseasonal forecasts up to 42 days ahead. It would also be interesting to extend the testing data, though as we can see in Supplementary Figure S11 it would not change much forecast skill global metrics such as ACC. This suggests sufficient stability for global metrics (with of course some possible not-statistically-significant differences in Rt maps at the small regional scale) and thus that overall results would qualitatively be robust if extended. Equally interesting would be the extension of the hindcasts years so as to have a much more reliable climatology for each considered model (Cf. Section Methods, subsection Climatology). However, the limitation of our set-up does not allow us to train and store more years of data.

While effective, our fine-tuning approach has two inherent limitations: it directly predicts weekly means and requires distinct model variants for different lead-times. These limitations restrict the application of our model to, for example, streamflow forecasts[31] and prediction of the MJO[13], which require higher temporal resolution forecasts to capture daily propagation. To address these limitations, future works could take advantage of our computationally cost-efficient fine-tuning method to make several model variants with a seamless evolution of the temporal resolution progressively decreasing with lead-times: e.g. from 6-hourly means for lead-times up to about 5-day lead, to daily means for about 5-10 day lead-times, to 2-3 day means for about 10-20 day leads, to weekly means for 3-5 week leads, to biweekly means for 6-10 week leads, to monthly means for seasonal forecasts, etc… Another path would be to develop fine-tuning methods optimized for auto-regressive rollouts and daily-scale predictions. Such an approach would naturally complement our ensemble generation framework, enabling multiple FT iterations to produce ensemble forecasts without lead-time specific training.

Another possible improvement is the use of other base models for fine-tuning, of higher spatial resolution, such as the 0.25° version of Graphcast or its equivalent at ECMWF, AIFS[32], so as to test our strategy on other systems and to obtain subseasonal forecasts at a higher spatial resolution. These should better capture the small-spatial scale processes such as orographic features, e.g. in mountain ranges such as the Himalayas, and in high islands such as Tahiti in French Polynesia or Big Island in Hawaii archipelagos. The implementation is straight-forward, but obviously needs a larger set of higher-resolution reanalysis data and more memory to retrain the model.

Another possible path for improvement would involve applying the fine-tuning method to GenCast[33], the inherently probabilistic variant of GraphCast. This approach might be more efficient, as it eliminates the need for external ensemble modeling techniques but the trade-off is that such probabilistic diffusion models are more computationnaly-costly. GenCast's foundation in diffusion models means each forecast is a discrete sample from the underlying generative process, directly providing a probabilistic prediction distribution. Fine-tuning on ERA5 data could therefore calibrate this entire distribution for subseasonal timescales, enhancing the reliability and spread of its ensemble members.

The natural next step would be to implement a 'OceAtmFT' model trained on both ocean and atmosphere variables. Indeed, integrating oceanic predicted variables, such as SST and Sea Level Anomaly (SLA), into FT could be an effective approach for enhancing its forecasting skill. Indeed, slow changing SST conditions provide a valuable source of predictability at subseasonal time-scales[34,35]. Also, integrating land moisture and ice-snow variables can be seen as a relevant way to improve subseasonal forecasts over land. FengShun-CSM[14] implemented such a model and achieved a better performance than the reforecasts of S2S ensemble mean.

In summary, the subseasonal predictions of a cutting-edge medium-range weather forecasting model (here GraphCast) have been enhanced both by our fine-tuning and ensemble forecasting techniques. In fact, the FT ensemble model surpasses significantly the

traditional dynamical S2S ensemble model for week 3 and 4 lead-times globally. The computation-cost effectiveness of the process is taken into account by both of those techniques, which allow for the production of ensemble predictions, which can be prohibitively expensive for dynamical models. Ensemble forecasts that incorporate ocean variables and other land moisture and ice-snow variables offer the most promising opportunity to further enhance subseasonal forecasting accuracy. This approach would furthermore provide a more comprehensive understanding of the earth ocean-atmosphere system. This ocean-atmosphere coupling represents the most relevant subsequent step in the development of subseasonal forecasts that are even more precise.

Methods

Data

ERA5

ERA5 is the fifth iteration of the ECMWF reanalysis dataset and offers a rich array of surface and upper-air variables. It operates at an approximately 25-km horizontal resolution (0.25°) and a temporal resolution of 1 hour, spanning from 1940 to the present day[16]. ERA5 stands as the most comprehensive and precise reanalysis archive globally and is considered the best-known estimation for most atmospheric variables[36,37]. In this study, we utilize 6-hourly ERA5 reanalysis, regridded to a 1° resolution, to fine-tune Graphcast. This data was obtained through the WeatherBench2[10] python package, a publicly available, cloud-optimized ground truth and baseline datasets. It serves as the sole data for fine-tuning Graphcast.

S2S

The S2S prediction project[38] was initiated in 2013 from an international effort to improve and develop various aspects of dynamical subseasonal predictions, including tropical cyclones[39] (TCs). The S2S project has created an extensive dataset containing subseasonal forecasts and reforecasts (also known as hindcasts) from 11 operational and research centers. A key goal of these efforts is to improve the forecasting skill and understand the sources of subseasonal predictability, notably the MJO, with the best forecasts of the MJO originating from ECMWF[4].

Often, operational subseasonal forecasting models are updated by incorporating recent research discoveries optimized for operational use. For instance, the ECMWF S2S hindcasts are generated on-the-fly by employing the most recent model version available at the time of forecast generation.

This study utilizes the hindcasts from ECMWF S2S CY49R1 model cycle with a resolution of 1.5°, as the reference dynamical model for the forecast verification period of 3 years from 2020 to 2022. Specifically, the verification process utilizes probabilistic forecasts, which contain 11 members. The S2S ensemble is also a coupled ocean-atmosphere model with full

ocean physics. This includes the explicit simulation of equatorial Kelvin and Rossby waves and oceanic mixed layer thermodynamics, which are critical for accurately forecasting the evolution of SST anomalies in the tropics.

Models

Graphcast Operational

The Graphcast medium-range model produces accurate weather forecasts at a resolution of 0.25° and has been evaluated up to 10 days ahead against the top deterministic operational system in the world, the ECMWF's high-resolution forecast (HRES). HRES has a 9 km horizontal resolution and is a product of the Integrated Forecasting System (IFS) that produces a 10-day forecast. After regridding to a resolution of 0.25°, Graphcast significantly outperforms HRES on 90% of 1380 verification targets[9].

As stated previously, this particular model takes as input two ERA5 reanalysis 6-hourly mean datasets, i.e the current time (usually at 00:00:00 UTC) and 6 hours earlier and forecasts the next atmospheric state 6 hours ahead. Graphcast is also an autoregressive model meaning that it can be "rolled-out" by feeding its own predictions back in as input, to generate weather forecasts at long lead-times (multiples of 6-hr).

The operational Graphcast was trained with 39 years of ERA5 reanalysis data from 1979 to 2017. As stated previously, it took roughly 4 weeks to train the model, on a 32 Go Cloud TPU and now can make accurate predictions in under a minute on a single TPU. Graphcast is based on the Graph Neural Network (GNN) in an "encoder-decoder" configuration and a has a total number of 36.7 million of parameters (demo available at https://github.com/deepmind/graphcast).

Training workflow for fine-tuning Graphcast and develop FT

The present study presents a strategy for fine-tuning the medium-range Graphcast global weather forecasting model to enhance its predictions at longer lead-times, on a subseasonal time-scale, combined with an ensemble forecasting technique to produce probabilistic forecasts from 11 different versions of FT. The FT ensemble model has three 'meta-versions', one for each lead time, and provides weekly global forecasts with a resolution of 1° for lead-times of 7-14, 14-21 and 21-28 days ahead. The forecasts generated by FT ensemble incorporate, as GraphCast, 6 upper-air atmospheric variables at 13 pressure levels and 5 surface variables.

Given the constraints of our local computational resources, we opted to leverage a Google Cloud environment, which grants us access to A100 Graphics Processing Units (GPUs). This allowed us to execute the training, inference, and fine-tuning procedures of the "small" Graphcast model. This "small" model is a version of Graphcast, with a resolution of 1° and reduced computational demands. However, it still includes a wide range of variables, incorporating 6 upper-air atmospheric variables at 13 pressure levels and 5 surface variables.

This "small" Graphcast was trained using ERA5 data from 1979 to 2015 and fine-tuned from 2016 to 2019. The training and fine-tuning period were chosen so as to not overlap with each other.

The Graphcast model requires, as aforementioned, two instantaneous ERA5 reanalysis as input, which are 6 hours apart. It then generates a forecast for the ERA5 reanalysis dataset 6 hours ahead. This process has been altered throughout the fine-tuning phase with the precise goal of training Graphcast to excel in subseasonal weekly forecasts. To do this, we followed the procedure outlined in Supplementary Figure S13.

In other words, the input of the original Graphcast model was composed of two atmospheric states, the most recent time step before the forecast starts and the previous state separated by a 6-hr timestep. Graphcast aimed to forecast the next 6-hr timestep. While FT input uses as previous state the previous week's state and aims to forecast weekly values directly for week 2, 3 and 4 lead-times.

Then 3 Graphcast models are fine-tuned separately with the custom data for week 2, week 3 and week 4 on 20 epochs, with a learning rate of 1e-4, without checkpoints, with an adam optimizer[40] and batch size = 1. The fine-tuning is first carried out on years 2016 to 2019.

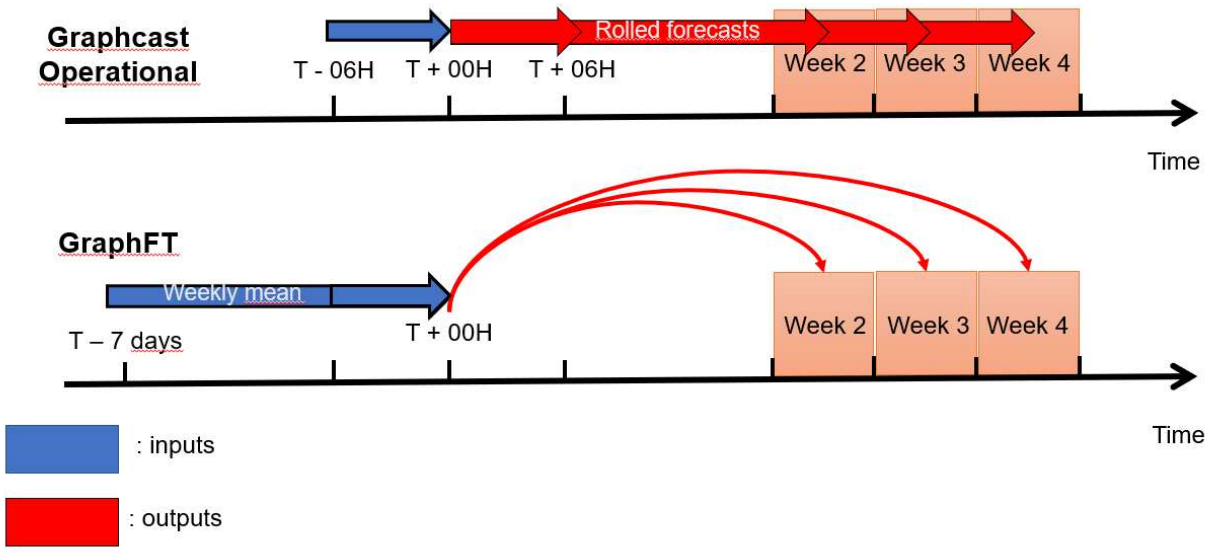


Figure 4. Schematic of the fine-tuning strategy aimed at improving intraseasonal forecasts, showing the differences in inputs and outputs between REF (upper) and FT (lower).

.

Following the process of fine-tuning, the accuracy of FT's predictions is assessed by comparing them to the forecasts generated by Graphcast Operational and the S2S ensemble dynamical model. The metrics utilized for the evaluation process are described in the Supplementary Materials.

Implementation of ensemble forecasting with FT

In order to implement ensemble forecasting using the fine-tuning of GraphCast, we utilized a novel method consisting of using different versions of FT as member. The difference from those versions being the years used as training. In the following table is presented all of the versions used for ensemble forecasting. 4 FTs were trained separately on each year between 2016 and 2019, 4 FT were trained on different 2-year combinations, and 2 FT have more than 2 years training. With the FT trained on 2016-2017-2018-2019 being the 'control version', the deterministic model of FT.

Table 1. Versions of FT used for ensemble modelling

Versions with individual years	Version with 2 years	More than 2 years
2016, 2017, 2018, 2019	2016-2018, 2017-2018, 2016-2019, 2018-2019, 2017-2019	**2016-2017-2018-2019***, 2017-2018-2019

* : control version with the longest training period, used for deterministic forecasts, refered to as 'FT'

Climatology

To remove the seasonal cycle from our forecasts and for subsequent model evaluation, two complementary methods were tested. First a daily climatology was derived from the ERA5 reanalysis dataset (Copernicus Climate Change Service, ECMWF) over the 1979–2019 period, providing a baseline for estimating the seasonal cycle. The same dataset, covering the period 2020–2022, serves as observational ground truth and is used to assess the accuracy of all forecasting models.

Usually, in operational forecasts, the seasonal cycle is removed by generating a set of hindcasts for each model, from which a model-specific climatology is computed and subtracted from the corresponding forecasts. For instance, the FuXi-S2S[13] model employs 14 years of hindcasts for this purpose. In principle, we aimed to follow this established approach. However, practical constraints arose due to the substantial data storage requirements. For example, storing three years of S2S ensemble forecasts already amounted to approximately 400 GB. Producing and storing a decadal hindcast climatology for each model under consideration exceeded our available computational and storage resources.

The Figure S14, in supplementary materials, compares two climatological approaches: the first row displays the temporal correlation coefficient (R_t) and root-mean-square error

(RMSE) for all models processed using the ERA5 climatology for T2M; the second row shows results obtained using a climatology derived from each model's own hindcasts still for T2M. These hindcasts were generated over the test period (2020–2022), rather than a separate historical period.

In terms of the Rt correlation coefficient, no substantial differences are observed between the two methods, with the exception of a slight degradation in accuracy for the deterministic FT model.

In contrast, a marked reduction in RMSE is evident when using the model-specific hindcast climatology. For instance, the RMSE for the FT ensemble at week 4 decreases from approximately 1.7 K to 1.2 K. With this approach, the deterministic FT model even slightly outperforms the S2S ensemble at the same lead time.

Even though there are minimal differences in correlation and such a significant increase in RMSE across models, we decided to adopt the climatology derived from ERA5 reanalysis, given the fact that there are not enough forecasts for each model to implement a proper climatology, and that we wanted to have proper hindcasts. Therefore, we expect the FT ensemble skill to further increase in future studies with more computational capacities.

Acknowledgements.

This work was mainly funded by the IRD (the French national Institute of Research for sustainable Development) through a 2-year post-doctoral 'overseas' funding of its 'OCEANS' department, and secondly through a 2-year post-doctoral funding from the PPR IA Ocean (as the present study establishes a good conceptual base for also improving through fine-tuning ocean forecasting models). The work was carried out at the UMR 241 SECOPOL (ex-EIO) at the University of French Polynesia (UPF), in collaboration with the French national weather service, Météo-France, and with the GePaSud UPF laboratory. We therefore thank IRD for its continuous support of our research, as well as the PPR IA Ocean, Météo-France, the UPF, UMR241 SECOPOL and GePaSud and all of our colleagues.

Data availability statement

Data and code are provided with this article. The ECMWF S2S data were obtained from https://apps.ecmwf.int/datasets/data/s2s/. For training and testing FT, we downloaded a subset of the ERA5 dataset obtained through the WeatherBench2[10] python package and that can be also obtained from https://cds.climate.copernicus.eu/, the official website of Copernicus Climate Data (CDS). Our models were trained in a Google Colab environment.

Competing interests

No competing interests

Author contributions

Conceptualization, V.S., T.I., M.H.; methodology, V.S., T.I., M.H., D.S.; software, V.S., T.I., M.H.; validation, V.S., T.I., M.H., D.S.,S.M-L. ; formal analysis, V.S., T.I.; investigation, V.S., T.I.; resources, V.S., M.H. ; data curation, V.S. ; writing—original draft preparation, V.S.; writing—review and editing, V.S., T.I., M.H., D.S, S.M-L.; supervision, T.I., M.H.; project administration, T.I., M.H.; funding acquisition, V.S.,T.I., M.H.,S.M-L.; All authors have read and agreed to the published version of the manuscript.

1. *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*. (National Academies Press, Washington, D.C., 2016). doi:10.17226/21873.
2. Richter, J. H. *et al.* Quantifying sources of subseasonal prediction skill in CESM2. *Npj Clim. Atmospheric Sci.* **7**, 1–9 (2024).
3. Coughlan de Perez, E. *et al.* Action-based flood forecasting for triggering humanitarian action. *Hydrol. Earth Syst. Sci.* **20**, 3549–3560 (2016).
4. Vitart, F. & Robertson, A. W. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Npj Clim. Atmospheric Sci.* **1**, 1–7 (2018).
5. White, C. J. *et al.* Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorol. Appl.* **24**, 315–325 (2017).
6. Chen, M., Wang, W. & Kumar, A. Prediction of Monthly-Mean Temperature: The Roles of Atmospheric and Land Initial Conditions and Sea Surface Temperature. https://doi.org/10.1175/2009JCLI3090.1 (2010) doi:10.1175/2009JCLI3090.1.

7. Doblas, F., García-Serrano, J., Lienert, F., Biescas, A. & Rodrigues, L. Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev. Clim. Change* **4**, (2013).
8. Vitart, F., Robertson, A. & Anderson, D. Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. *WMO Bull.* **61**, (2012).
9. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
10. Rasp, S. *et al.* WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024).
11. Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
12. Pathak, J. *et al.* FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. Preprint at http://arxiv.org/abs/2202.11214 (2022).
13. Chen, L. *et al.* A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.* **15**, 6425 (2024).
14. Zhou, C. *et al.* A machine learning model for skillful climate system prediction. *arXiv.org* https://arxiv.org/abs/2505.06269v1 (2025).
15. Kochkov, D. *et al.* Neural general circulation models for weather and climate. *Nature* 1–7 (2024) doi:10.1038/s41586-024-07744-y.
16. Hersbach, H. *et al.* The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

17. Ling, F. *et al.* FengWu-W2S: A deep learning model for seamless weather-to-subseasonal forecast of global atmosphere. *arXiv.org* https://arxiv.org/abs/2411.10191v2 (2024).
18. Cresswell-Clay, N. *et al.* A Deep Learning Earth System Model for Efficient Simulation of the Observed Climate. *AGU Adv.* **6**, e2025AV001706 (2025).
19. Wang, C. *et al.* Coupled Ocean-Atmosphere Dynamics in a Machine Learning Earth System Model. Preprint at https://doi.org/10.48550/arXiv.2406.08632 (2024).
20. He, S., Li, X., DelSole, T., Ravikumar, P. & Banerjee, A. Sub-Seasonal Climate Forecasting via Machine Learning: Challenges, Analysis, and Advances. *Proc. AAAI Conf. Artif. Intell.* **35**, 169–177 (2021).
21. Kiefer, S. M., Lerch, S., Ludwig, P. & Pinto, J. G. Can Machine Learning Models Be a Suitable Tool for Predicting Central European Cold Winter Weather on Subseasonal to Seasonal Time Scales? https://doi.org/10.1175/AIES-D-23-0020.1 (2023) doi:10.1175/AIES-D-23-0020.1.
22. Chen, L. *et al.* FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *Npj Clim. Atmospheric Sci.* **6**, 1–11 (2023).
23. Domeisen, D. I. V. *et al.* Advances in the Subseasonal Prediction of Extreme Events: Relevant Case Studies across the Globe. https://doi.org/10.1175/BAMS-D-20-0221.1 (2022) doi:10.1175/BAMS-D-20-0221.1.
24. de Andrade, F. M., Coelho, C. A. S. & Cavalcanti, I. F. A. Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Clim. Dyn.* **52**, 5451–5475 (2019).

25. Christidis, N. & Stott, P. A. Changes in the geopotential height at 500 hPa under the influence of external climatic forcings. *Geophys. Res. Lett.* **42**, 10,798-10,806 (2015).
26. Chen, T., Xu, B., Zhang, C. & Guestrin, C. Training Deep Nets with Sublinear Memory Cost. Preprint at https://doi.org/10.48550/arXiv.1604.06174 (2016).
27. Evensen, G. The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation. *Ocean Dyn.* **53**, 343–367 (2003).
28. Lang, S. *et al.* AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *Npj Artif. Intell.* **2**, 18 (2026).
29. Wheeler, M. C. & Hendon, H. H. An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. *Mon. Weather Rev.* **132**, 1917–1932 (2004).
30. Rocuet, C. *et al.* Key role of the Madden-Julian Oscillation on humid heatwaves. Preprint at https://doi.org/10.48550/arXiv.2509.17526 (2025).
31. Nguyen, D. H. *et al.* Advancing Sub-Seasonal to Seasonal Streamflow Forecasting in Canada: A Review of Conventional and Emerging Approaches for Operational Applications. *Results Eng.* **27**, 106345 (2025).
32. Lang, S. *et al. AIFS - ECMWF's Data-Driven Forecasting System*. (2024).
33. Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
34. Albers, J. R. & Newman, M. Subseasonal predictability of the North Atlantic Oscillation. *Environ. Res. Lett.* **16**, 044024 (2021).

35. Yan, Y., Liu, B. & Zhu, C. Subseasonal Predictability of South China Sea Summer Monsoon Onset With the ECMWF S2S Forecasting System. *Geophys. Res. Lett.* **48**, e2021GL095943 (2021).

36. Linus Magnusson, S. M. Tropical cyclone activities at ECMWF. *ECMWF* https://www.ecmwf.int/en/elibrary/81277-tropical-cyclone-activities-ecmwf (2021).

37. Jiao, D., Xu, N., Yang, F. & Xu, K. Evaluation of spatial-temporal variation performance of ERA5 precipitation data in China. *Sci. Rep.* **11**, 17956 (2021).

38. Vitart, F. *et al.* The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bull. Am. Meteorol. Soc.* **98**, 163–173 (2017).

39. Lee, C.-Y., Camargo, S. J., Vitart, F., Sobel, A. H. & Tippett, M. K. Subseasonal Tropical Cyclone Genesis Prediction and MJO in the S2S Dataset. *Weather Forecast.* **33**, 967–988 (2018).

40. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv.org* https://arxiv.org/abs/1412.6980v9 (2014).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterialsnpjTIVS.docx
- graphcastfinetuningscript.zip