MODELLING SLOW AND FAST VARIABLES FOR SUBSEASONAL-TO-SEASONAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models for weather prediction have reached impressive accuracy. Yet, a significant performance and reliability gap remains at subseasonal-to-seasonal (S2S) forecasting horizons, i.e., 15 days to 2 months. To reach high predictive skill in this regime, accurate estimates are required not only for fast atmospheric dynamics but also for slowly changing boundary conditions on the Earth's surface (for instance, sea surface temperature). This makes S2S dynamics a coupled multi-scale system in which not all variables evolve at the same pace. We address this challenge by extending the Earth System Foundation Model (ESFM), a transformer-based model of weather and climate, such that it is able to explicitly handle the different update frequencies. We introduce a Temporal Attention Aggregator (TAA) taking as input samples from arbitrary past timesteps, and devise a mixed-lead-time (mixed-LT) approach encouraging the model to focus on selected variables' timesteps for each time scale. Our extended TAA model for S2S forecasting (TAAM-S2S), trained with adapted block rollout finetuning, improves the skill of the standard ESFM baseline for *slow* variables, while matching its accuracy for *fast* variables at S2S time scales. Moreover, TAAM-S2S outperforms DLESyM, AIFS v1.1 and Aurora v1.5 in terms of ensemble metrics for *slow* variables. It also reaches competitive performance in a rare-event case study, predicting droughts in Sichuan in 2022 and in Amazonia in 2023.

1 INTRODUCTION

The rapid developments in deep learning in recent years have brought some undisputed successes in the field of weather forecasting. Data-driven models have surpassed the accuracy of numerical weather predictions (NWP) on standard metrics, while producing forecasts at a fraction of the cost of their predecessors (Bi et al., 2023; Lam et al., 2023; Lang et al., 2024; Bodnar et al., 2025; Bonev et al., 2025; Ozdemir et al., 2026). However, despite deep learning models thriving at mid-range forecasts, their skill over extended horizons is only beginning to catch up with NWP (Loegel et al., 2025), facing challenges such as error accumulation and blurring over long autoregressive rollouts.

Forecasts at the subseasonal-to-seasonal (S2S) timescale, which conventionally spans from two weeks to two months, are increasingly vital for natural disaster risk planning, agricultural management and renewable energy scheduling (White et al., 2017). Unlike mid-range forecasts, which are largely governed by atmospheric dynamics and initial conditions, S2S predictive skill stems additionally from slowly evolving boundary conditions at the Earth's surface and large-scale climate modes. Recent literature demonstrates that Earth system components such as sea surface temperature, sea-ice cover and soil moisture may provide memory and feedback mechanisms that drive predictive skill across extended lead times (Tokuda & Dirmeyer, 2026; Cresswell-Clay et al., 2025). Therefore, effectively encoding slowly varying land-surface and oceanic boundary variables is a promising direction for further improvements at S2S timescales.

However, it is challenging to model variables with differing temporal dynamics in a single model. While the common 6-hour lead time is effective for atmospheric variables, the evolution of slow variables at this time step is dominated by noise. Hence, to properly model the Earth system evolution, each variable should be trained and unrolled at its own pace. Although the predominant paradigm is training a model for a single fixed autoregressive step size, some AI models allow for various lead times, either by training lead time-specific sub-models (Bi et al., 2023; Chen et al., 2024) or by con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ditioning inputs on the lead time (Nguyen et al., 2023; Andrychowicz et al., 2023). However, this approach has been proposed mainly to reduce error accumulation and does not account for different temporal dynamics. On S2S timescales specifically, the lead time is usually chosen larger than in weather forecasting, typically from one day to one week (Chen et al., 2024; Li et al., 2026).

By contrast, we propose one model that performs well on medium-range forecasts and excels at the S2S horizon. We aim for it to be stable in month-long rollouts while still being able to produce sub-daily predictions. Hence, we introduce the **Temporal Attention Aggregator (TAA)** submodule and apply it to the Earth System Foundation Model (ESFM) (Ozdemir et al., 2026) to learn the most effective past time steps for each lead time and variable categorized as *slow* or *fast* (Section 3.2). To maximize its effectiveness, we develop a **mixed-lead-time (mixed-LT)** training and inference strategy that predicts variables at various lead times while weighting the loss contribution of each variable depending on its signal-to-noise ratio and predictability (Section 3.3). Furthermore, we propose **block rollout finetuning**, an adaptation of the known mechanism that makes it suitable for the mixed-LT approach (Section 3.4). Our proposed model, TAA for S2S forecasting (TAAM-S2S), outscores state-of-the-art deep learning models in predicting sea surface temperature, sea-ice cover, and soil moisture, without losing the S2S performance on dynamic atmospheric variables (Section 5).

2 Related work

Deep Learning Weather Prediction (DLWP) has reached a maturity where many models outperform traditional NWP on standard metrics. While many DLWP models build on transformer architectures, (e.g., Pangu-Weather (Bi et al., 2023), FuXi (Chen et al., 2023), and Aurora (Bodnar et al., 2025)), models based on Graph Neural Network architectures have also demonstrated strong performance (e.g., GraphCast (Lam et al., 2023) and FastNet (Dunstan et al., 2026)) and are commonly used by operational weather services (AIFS (Lang et al., 2024)). These models are usually evaluated up to 15 days, the time horizon beyond which their predictions tend to become blurry or collapse to climatology (Decancq et al., 2026; Ben Bouallègue et al., 2024), leaving an unresolved gap to subseasonal predictions.

Subseasonal-to-seasonal (S2S) forecasts have traditionally been addressed with numerical models that extend NWP to longer horizons, leading to large comparison databases such as the S2S database (Vitart et al., 2017). Machine Learning methods have been employed early on to postprocess these simulations and increase their accuracy (Scheuerer et al., 2020; Horat & Lerch, 2024). They remain highly accurate S2S methods in recent benchmarks such as the WeatherQuest challenge (Loegel et al., 2025), along with hybrid methods that combine a numerical core with learned components (Kim et al., 2025; Guan et al., 2026).

Due to the complexity of S2S forecasts, purely data-driven models specializing in this field have mostly focused on predicting aggregated outputs, such as biweekly averages at specific locations (Hwang et al., 2019; He et al., 2021), weekly-averaged global maps (Lyu et al., 2026; Sansine et al., 2025; Zhou et al., 2026) and specialized climate modes (Delaunay & Christensen, 2022). A precursor work by Weyn et al. (2021) highlighted the promises of AI for S2S predictions on a handful of variables. Only recently have global S2S AI models with a fully resolved atmosphere emerged. They inherit from architectures developed for weather forecasting and extend them with probabilistic components, such as Fuxi-S2S (Chen et al., 2024), AIFS-SUBS (Schloer et al., 2026), and TianQuan-S2S (Li et al., 2026). However, these models predict daily averages that neglect the diurnal cycle and have not investigated the potential benefits of a multi-scale approach. Initial multi-scale approaches focused on coupling atmospheric and oceanic dynamics and studied S2S as a by-product of their development (DLESyM (Cresswell-Clay et al., 2025) and SamudrACE (Duncan et al., 2026)).

3 Methodology

3.1 Earth System Foundation Model (ESFM)

The ESFM is a vision-transformer-based foundation model for the Earth system, flexible to various input types (Ozdemir et al., 2026) whose core components inherit from Aurora (Bodnar et al., 2025).

Like most AI weather models, ESFM takes as inputs two previous states of the Earth at time $t_0 - \delta$ and t_0 to produce the output for $t_0 + \delta$ and is applied autoregressively to obtain consecutive predictions. The ESFM encoder encodes each physical variable individually from feature map patches to vectors. Then, it computes axial attention between the tokens and aggregates them with a Perceiver module (Jaegle et al., 2021). We apply the Temporal Attention Aggregator (TAA), described in Section 3.2, in place of variable-specific patch embeddings, leveraging separate processing flows for each variable, in particular for variables evolving at different paces.

ESFM consists of $\sim$110 million trainable parameters. Its processor is a 3D Swin Transformer (Liu et al., 2021) with four latent channels, three coming from the atmospheric variables and one from the surface variables. We modify this routing by assigning the latent channels to different temporal dynamics: the atmospheric variables are limited to two latent channels, one is dedicated to fast surface variables, and one to slow surface variables (upper ocean and soil moisture).

3.2 Temporal Attention Aggregator (TAA)

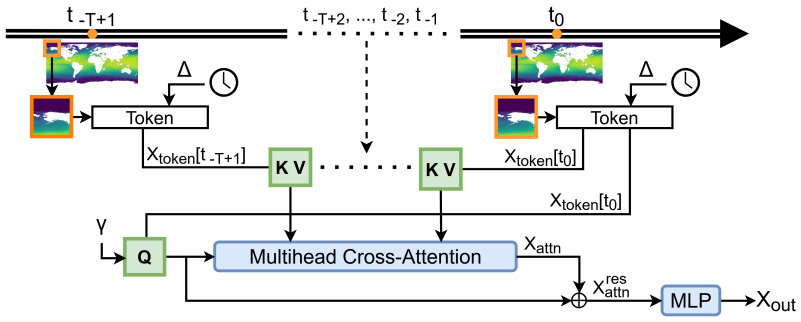

Figure 1: Temporal Attention Aggregator (TAA), applied independently to each surface variable and spatial patch. Patches from all T input timesteps are encoded with the ESFM patch embedding and enriched using relative time embedding Δ. The t_0 token augmented with the elapsed-time embedding γ forms the query. All T tokens form keys and values. The attention output is added residually to the query token and passed through an MLP.

Slow variables require a longer input history than the two timesteps traditionally used by AI weather models. We introduce the Temporal Attention Aggregator (TAA), which replaces the variable-specific patch embedding of surface variables to encode T input timesteps $(t_{-T+1}, \ldots, t_{-1}, t_0)$ at a cost linear in T, without increasing the number of tokens entering the processor. Note that the timesteps need not be uniformly spaced, and that in this work, each tasked lead time corresponds to the input timestep ($t_0 - t_{-i} = t_i - t_0 = \delta_i$). Since the most informative past timesteps are not known a priori and may differ across variables and lead times, TAA learns them with cross-attention. As illustrated in Figure 1, each variable at each timestep is encoded into its token embedding using the original ESFM patch embedding layer. Then, we add the time embedding $\Delta(\delta_i)$ of its offset relative to t_0:

$$X_{token}[t_{-i}] = X_{embed}[t_{-i}] + \Delta(\delta_i) \tag{1}$$

Attention operates along the temporal axis, independently for each variable and spatial patch. All T tokens produce the keys and values, while the most recent token $X_{token}[t_0]$, enriched with the embedding $\gamma(\delta_j)$ of the prediction lead time δ_j, produces the single query. TAA thus replaces the two-timestep patch embedding of (t_{-1}, t_0) with a transformer block:

$$\begin{aligned} X_{attn} &= \mathrm{MHA}(X_{token}[t_0] + \gamma(\delta_j), [X_{token}[t_0], \ldots, X_{token}[t_{-T+1}]]) \,, \\ X^{res}_{attn} &= \mathrm{LN}_1(X_{attn} + X_{token}[t_0] + \gamma(\delta_j)) \,, \\ X_{out} &= \mathrm{LN}_2(X^{res}_{attn} + \mathrm{MLP}(X^{res}_{attn})) \,, \end{aligned} \tag{2}$$

where LN is layer norm and $\mathrm{MHA}(q, KV)$ is multi-head attention with query q and key/value sequence KV. Note that the query contains $\gamma(\delta_j)$ to allow the same variable to attend to different past timesteps depending on the forecast horizon.

Since TAA is applied separately to each variable, different variables can attend to different timesteps, making the separation of temporal dynamics most effective. We hypothesize that attention should

focus mostly on the two most recent timesteps for fast variables (e.g., wind speed), whereas slowly evolving variables should spread their attention over the past data (Section 4.2).

3.3 MIXED-LEAD-TIME (MIXED-LT) TRAINING AND INFERENCE

To handle multiple time scales, we train a single model with multiple lead times, taking advantage of the TAA attending to the full range of inputs. We call this approach mixed-lead-time (mixed-LT) training (Figure 2). Unless stated otherwise, the inputs are $(t_{-2},\ t_{-1},\ t_0)$. All variables are predicted at $\delta_1 = 6$h and $\delta_2 = 48$h lead times. The choice of lead times δ_1, δ_2 is investigated in Section 4.3 and results with five input timesteps are presented in Appendix G. The lead time is sampled randomly for each batch with a 6:4 ratio (δ_1:δ_2). Loss weights per variable type and lead time give equal contribution to all variables at δ_1 ($w_{\delta_1}^{fast} = 1$, $w_{\delta_1}^{slow} = 1$) and emphasize slow variables at δ_2 ($w_{\delta_2}^{fast} = 0.01$, $w_{\delta_2}^{slow} = 1$).

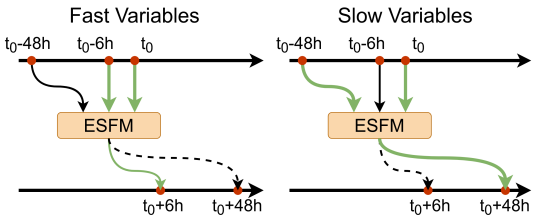

Figure 2: Mixed-lead-time (mixed-LT) scheme, with the model focusing on fast variables for short δ_1 and slow variables for distant δ_2.

The mixed-LT procedure also requires special handling at inference time to benefit from the model accuracy on both slow and fast variables. The first δ_2 prediction gives the slow variables at $t_0 + \delta_2$, which are linearly interpolated to $t_0 + \delta_1, t_0 + 2\delta_1, \ldots, t_0 + (n-1)\delta_1$, with $n = \delta_2/\delta_1 \in \mathbb{N}$. The fast variables are then rolled out autoregressively with step δ_1, using the interpolated slow variables as inputs. The system thus evolves at the finest temporal granularity, while reducing error accumulation and mitigating the low signal-to-noise ratio for slow variables at short lead times. This *inference block* is repeated for the following blocks of length δ_2.

3.4 BLOCK ROLLOUT FINETUNING

Rollout finetuning is commonly used in DLWP to reduce the distribution shift between ground-truth inputs during training and model predictions at inference (Lang et al., 2024; Dunstan et al., 2026). To limit the size of the computational graph, gradients are backpropagated only through the last rollout step thanks to the pushforward trick (Brandstetter et al., 2022). With mixed-LT, the pushforward trick must be adapted. The input $t_{-2} = t_0 - \delta_2$ reaches one block into the past, so the first two inference blocks still receive some ground-truth inputs. We therefore unroll three inference blocks and truncate the computation graph at the block level. For each block, we add the δ_2 loss to the mean of the n δ_1 losses, backpropagate, update the weights, and proceed to the next one.

To avoid the excessive smoothing induced by rollout finetuning, we follow Dunstan et al. (2026) and add a weighted Modified Spherical Harmonic (MSH) loss on the fast variables, obtaining $\mathcal{L}_{rollout} = \mathcal{L}_{MAE} + w_{MSH}\mathcal{L}_{MSH}$. It penalizes spectral errors with a wavenumber-dependent weight to emphasize small scales (Appendix A). Its effect is reported in Section 4.5.

3.5 DATA

We use the ERA5 (Hersbach et al., 2020) at 1.5° x 1.5° resolution with 6-hourly timesteps (00, 06, 12, 18), as curated by Rasp et al. (2024) using conservative regridding. We train on 1979-2018, validate on 2019, and hold out 2020 onwards for testing. The atmosphere is represented by five variables (geopotential Z, wind components U and V, temperature T, and specific humidity Q) on three pressure levels (1000 hPa, 850 hPa, and 500 hPa). This subset increases data loading speed and emphasizes the surface, where the slow variables are predicted, at the cost of omitting stratospheric levels known to be subseasonal predictors. *Fast* surface variables are 2-meter temperature (2T), 10-meter wind components (10U, 10V), mean sea-level pressure (MSL) and 6-hour accumulated precipitation (TP6h). *Slow* variables are sea surface temperature (SST), sea-ice cover (CI) and volumetric soil water in four layers (SWVL1-4). Details are given in Appendix B.

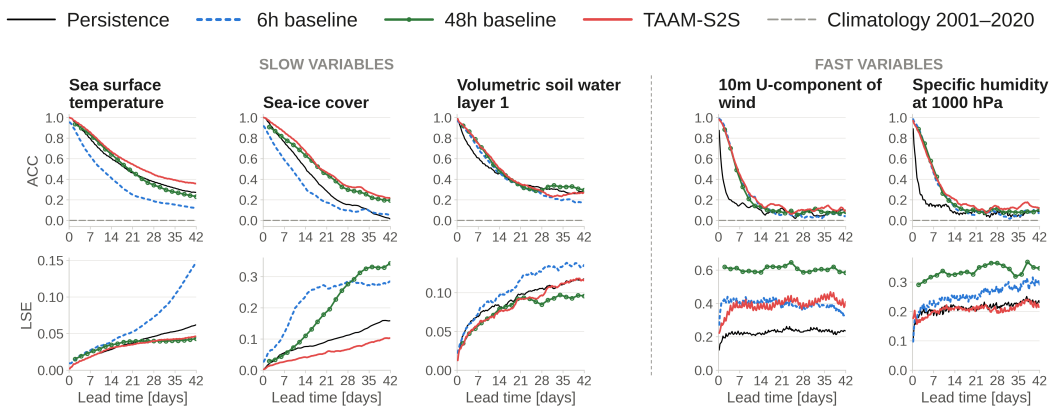

Figure 3: ACC (↑) and LSE (↓) metrics as a function of lead time for selected slow and fast variables, scoring TAAM-S2S against ESFM-based baselines, persistence and climatology.

4 Mixed-lead-time models

4.1 Training procedure

All models are trained on one NVIDIA GH200 node with AdamW, and a cosine learning rate scheduler with linear warmup. Unless stated otherwise, we use a batch size of 4 and the latitude- and variable-weighted MAE loss (Appendix A). Training has three stages:

1. The 6h and 48h ESFM baselines are trained from scratch for 100 epochs using learning rate of 5e-4 and 1000-step warmup, taking as input $(t_0 - 6\text{h}, t_0)$ and $(t_0 - 48\text{h}, t_0)$, respectively.
2. TAA replaces the variable patch embedding, and encoder weights are reinitialized. The model is trained with mixed-LT for 10 epochs using learning rate of 5e-4 and 1000-step warmup. The lead time ablation starts from the 6h baseline (Section 4.3), and TAAM-S2S starts from the 48h residual baseline (Section 4.4). Longer trainings led to overfitting on some slow variables.
3. We finetune for 5000 steps with a batch size of one, learning rate of 5e-5 and 200 warmup steps (Section 4.5). w_{MSH} is set so the MAE and MSH loss terms are in 9:1, 7:3 or 5:5 ratios.

4.2 TAA model for S2S forecasting (TAAM-S2S)

Our TAA model for S2S forecasting (TAAM-S2S) is the model obtained after all three stages with the 5:5 loss ratio during block rollout finetuning. It uses three inputs (t_{-2}, t_{-1}, t_0) and two lead times $(\delta_1 = 6\text{h}, \delta_2 = 48\text{h})$. For slow variables, it predicts residuals instead of absolute values (Section 4.4). We compare TAAM-S2S to the 6h and 48h ESFM baselines and add climatology and persistence as references. Climatology is computed for each hour and day of the year as a mean for a given grid point over 20 years, from 2001 to 2020. Persistence is created by freezing the variable values from t_0. All metrics in this section are evaluated on 42-day rollouts started from 12 initial dates in 2021, one per month, to cover the whole annual cycle.

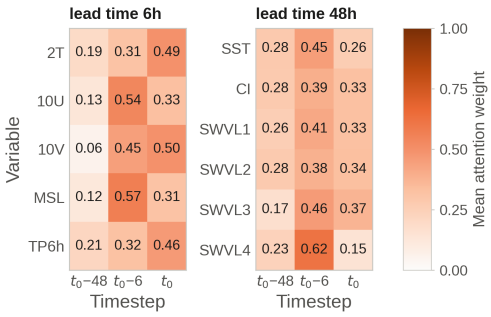

Figure 4: Attention weights extracted from the TAAM-S2S model.

We report three metrics, described in Appendix C. Root Mean Square Error (RMSE) measures pointwise accuracy. Anomaly Correlation Coefficient (ACC) captures the models' ability to forecast trends. Logarithmic Spectral Error (LSE) verifies that the predictions have the right amount of small-scale detail. TAAM-S2S gives superior performance to the 6h baseline and matches the 48h baseline in terms of RMSE and ACC on slow variables, often yielding improved LSE (see Figure 3).

Table 1: Mean and standard deviation of ACC (↑) for slow and fast variables calculated based on 12 S2S rollouts initialized in 2021. Best scores are highlighted in bold, second best are underlined.

Model	Slow variables (weeks 1-6)					
	SST	CI	SWVL1	SWVL2	SWVL3	SWVL4
6 h baseline	0.35 ± 0.07	0.29 ± 0.08	0.42 ± 0.07	0.52 ± 0.05	0.67 ± 0.04	0.84 ± 0.01
48 h baseline	0.51 ± 0.08	$\underline{0.48} \pm 0.10$	0.46 ± 0.08	0.60 ± 0.08	$\underline{0.79} \pm 0.03$	$\mathbf{0.96} \pm 0.00$
mixed-LT 6/48 h	$\underline{0.56} \pm 0.08$	0.41 ± 0.09	$\underline{0.49} \pm 0.07$	$\underline{0.62} \pm 0.07$	0.78 ± 0.04	0.92 ± 0.01
mixed-LT 6/96 h	0.50 ± 0.10	0.44 ± 0.08	0.45 ± 0.08	0.58 ± 0.07	0.77 ± 0.04	0.93 ± 0.00
mixed-LT 6/192 h	$\mathbf{0.62} \pm 0.07$	$\mathbf{0.53} \pm 0.09$	$\mathbf{0.50} \pm 0.08$	$\mathbf{0.64} \pm 0.06$	$\mathbf{0.81} \pm 0.03$	$\underline{0.95} \pm 0.00$
Model	Fast variables (weeks 1-2)					
	2T	TP6h	T850	Z500	V1000	Q1000
6 h baseline	$\mathbf{0.55} \pm 0.05$	$\mathbf{0.29} \pm 0.02$	$\mathbf{0.54} \pm 0.05$	$\mathbf{0.60} \pm 0.05$	$\mathbf{0.45} \pm 0.03$	$\mathbf{0.51} \pm 0.05$
48 h baseline	$\underline{0.50} \pm 0.07$	$\underline{0.26} \pm 0.02$	0.47 ± 0.07	0.52 ± 0.07	0.37 ± 0.05	0.46 ± 0.06
mixed-LT 6/48 h	0.40 ± 0.05	0.20 ± 0.01	$\underline{0.52} \pm 0.05$	$\underline{0.58} \pm 0.05$	$\underline{0.43} \pm 0.04$	$\underline{0.47} \pm 0.04$
mixed-LT 6/96 h	0.38 ± 0.09	0.20 ± 0.02	0.38 ± 0.07	0.42 ± 0.06	0.29 ± 0.04	0.36 ± 0.04
mixed-LT 6/192 h	0.38 ± 0.09	0.20 ± 0.02	0.40 ± 0.07	0.46 ± 0.05	0.31 ± 0.04	0.39 ± 0.04

It also scores consistently better in RMSE and ACC on fast variables, only achieving slightly worse spectral metrics for some of them. Results for the remaining variables are presented in Appendix E.

Figure 4 shows the TAA attention weights for one inference step at both lead times, averaged over the 12 initial dates. For fast variables at 6h lead time, the focus is put on the last two timesteps ($t_0 - 6\text{h}$, t_0). For slow variables at 48h lead time, the most distant timestep $t_0 - 48\text{h}$ receives more weight than for fast variables, which supports our hypothesis from Section 3.2.

4.3 Lead time selection

We study the influence of the δ_2 lead time selection, considering $\delta_2 \in \{48\text{h}, 96\text{h}, 192\text{h}\}$. We start from 48h, as this lead time was also used in other works to model ocean variables (Cresswell-Clay et al., 2025), and extended it to allow for an increased signal-to-noise ratio. These models are initialised from the 6h baseline, without residual predictions or block rollout finetuning. Table 1 reports the mean and standard deviation of ACC for weeks 1-6 for the slow variables and weeks 1-2 for the selected fast variables.

All mixed-LT models improve RMSE, ACC, and LSE across slow variables compared to the 6h baseline. The 6/48h model also keeps the same ACC for all 15 atmospheric variables, losing mainly on the dynamic surface variables. Overall, mixed-LT training with TAA combines the temporal resolution of the 6h baseline with slow-variable skill close to the 48h baseline, in one model trained for both tasks. Numerical results for other metrics are listed in Appendix F.

While 6/96h does not improve over 6/48h, extending δ_2 to 192h improves the slow variable skill further: the 6/192h model scores the best RMSE, ACC, and LSE across almost all slowly-evolving features. This comes at a cost on dynamic variables, whose ACC drops (Table 1) and power spectra degrade (Table 6). The model loses the ability to reconstruct details at small scales. The choice between 6/48h and 6/192h depends on the use case and on the trade-off between slow and fast variables performance. We use 6/48h for TAAM-S2S, as it preserves the skill on dynamic variables.

We also trained a single model to perform 6h, 48h, 96h, 192h and 384h lead time prediction at the same time. It learned all tasks with satisfactory results. However, for each lead time set as δ_2 during

Table 2: Mean and standard deviation of ACC (↑) for the mixed-LT 6/48h model predicting absolute feature maps and its equivalent optimised to predict residuals for slow variables.

Model	sea surface temperature		sea-ice cover		volumetric soil water layer 1	
	Weeks 1–2	Weeks 3–6	Weeks 1–2	Weeks 3–6	Weeks 1–2	Weeks 3–6
absolute	0.76 ± 0.05	0.46 ± 0.10	0.70 ± 0.08	0.27 ± 0.11	0.72 ± 0.06	$\mathbf{0.37} \pm 0.09$
residual	$\mathbf{0.85} \pm 0.03$	$\mathbf{0.47} \pm 0.07$	$\mathbf{0.84} \pm 0.03$	$\mathbf{0.40} \pm 0.12$	$\mathbf{0.75} \pm 0.06$	0.33 ± 0.09

rollout, it performed slightly worse than the corresponding model with only two lead times presented during training (Appendix G).

4.4 Residual predictions

For slow variables, we predict the change $\boldsymbol{X}_{t_i} - \boldsymbol{X}_{t_0}$ instead of the absolute feature maps $\boldsymbol{X}_{t_i}$. This makes it easier for the model to learn the direction of the signal in the presence of significant noise. Table 2 compares both targets for the mixed-LT 6/48h model initialised from the 48h baselines. Residual predictions improve ACC for all three variables, except volumetric soil water where the difference is within one standard deviation. Residual predictions also preserve the correct trend and its amplitude in multi-year rollouts (Appendix I).

When initialized from the 6h baseline, residual mixed-LT models learned to predict persistence for sea-ice cover, regardless of δ_2 (Figure 17). This is because $\sim 71\%$ of the ocean is never covered with ice, so predicting 0 there already reaches a low loss. Initialising from the 48h residual baseline, which is already good on this variable, overcomes this issue without losing any dynamic variable skill. Moreover, residual predictions for dynamic variables proved ineffective, as such models experienced too severe error accumulation (see Appendix H).

4.5 Block Rollout Finetuning

We ablate the MAE:MSH ratio used in the block rollout finetuning of TAAM-S2S (5:5) by comparing it with 9:1 and 7:3 variants, all starting from the same mixed-LT 6/48h residual model. Figure 5 shows the effect on the example of the 10m U-component of wind.

For all ratios, block rollout finetuning yields a similar significant improvement in the RMSE of dynamic variables at S2S lead times. The ratio mainly affects the power spectra, which are closer to the ERA5 ground truth when more emphasis is put on MSH. For all selected w_{MSH}, we improve the spectra for most wavelengths $<$ 2000 km, which corresponds to generating more variance for small-scale details. As an example, with the 5:5 ratio, the ratio of captured small-scale variance ($<$ 2000 km) of U10 improves from 42% to 60% for the first 6h step, and from 36% to 66% after 42 days of rollout. The costs of block rollout finetuning are slightly increased LSE for some slow variables.

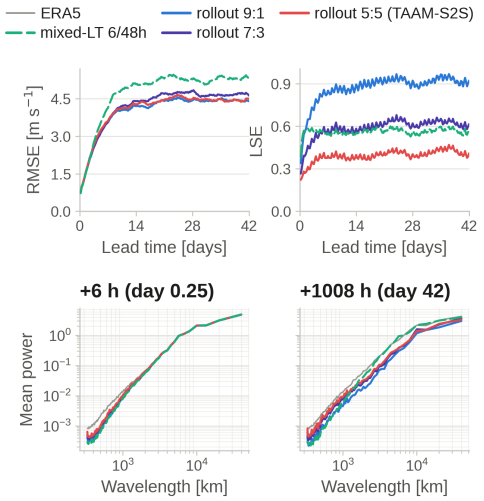

Figure 5: RMSE (↓), LSE (↓) and power spectra of 10m U-component of wind for the models after block rollout finetuning.

5 Benchmark against state-of-the-art models

5.1 S2S horizon forecasts

We compare TAAM-S2S against AIFS v1.1 (Moldovan et al., 2026), Aurora v1.5 (Weyn et al., 2026) and DLESyM (Cresswell-Clay et al., 2025). AIFS v1.1 contains all of our atmospheric and fast surface variables, as well as volumetric soil water layers 1-2. Aurora v1.5 predicts, among others, two of our slow variables: sea ice and volumetric soil water layer 1. DLESyM couples the atmosphere with the slowly evolving ocean, but sea surface temperature is its only slow variable, and it lacks most of our atmospheric features. Like TAAM-S2S, all models are used in the deterministic setting.

We assess all models on 12 rollouts of 42 days, one for each month of 2024. Benchmark models outputs are regridded to 1.5° x 1.5° with the same conservative procedure we used to create the ground truth dataset.

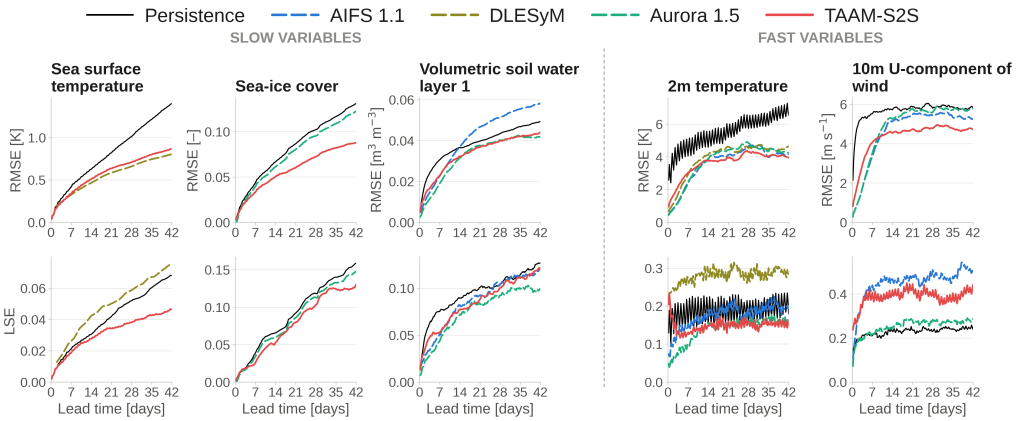

Figure 6: RMSE (↓) and LSE (↓) of the TAAM-S2S model, compared with AIFS v1.1, Aurora v1.5 and DLESyM, calculated on the 1.5° grid. Additional variables are given in Appendix K.

Figure 6 presents the results of deterministic forecasts. For dynamic variables, TAAM-S2S usually outperforms the other models in terms of RMSE at S2S lead times. For slow variables, TAAM-S2S roughly matches the RMSE of DLESyM on sea surface temperature and achieves a better LSE, despite DLESyM using a separate model dedicated to this variable. TAAM-S2S outperforms Aurora v1.5 on sea ice and matches it on volumetric soil water layer 1. Our model also scores better than AIFS v1.1 on both soil moisture layers. The downside of the mixed-LT approach is a higher LSE on some dynamic variables, mainly compared to Aurora v1.5.

Figure 7 shows the Ranked Probability Skill Scores (RPSS) for slow variables. They are computed from 12-member ensembles created by shifting back the initial time with 6h strides, binning each weekly variable mean into 2001-2020 climatology terciles. At longer lead times, TAAM-S2S outperforms all other models on sea surface temperature, sea-ice cover and soil water layer 2, and is close to Aurora v1.5 on layer 1. Beyond week 3, however, no model beats climatology on sea-ice cover and soil water layer 1.

5.2 Use case: drought prediction

We evaluate TAAM-S2S on drought forecasts, as those scenarios require capturing the interactions between the atmosphere and the soil, which we observe through the volumetric soil water layer outputs. The first event is the drought observed in the Yangtze river basin in the summer of 2022 (Lu et al., 2023). We focus on the Sichuan region (28.5-31.5°N, 105-108°E), one of the most severely impacted areas. We compare TAAM-S2S with Aurora v1.5 and AIFS v1.0, since AIFS v1.1 includes 2022 in its training data. The second event is the 2023 drought in Amazonia (15°S-4.5°N, 75-51°W) (Espinoza et al., 2024), where we compare against Aurora v1.5 and AIFS v1.1. For both events, we use the 12-member lagged ensembles, as in Section 5.1. The forecast initial date for Sichuan event is set to 1 July 2022 and for Amazonia to 1 August 2023.

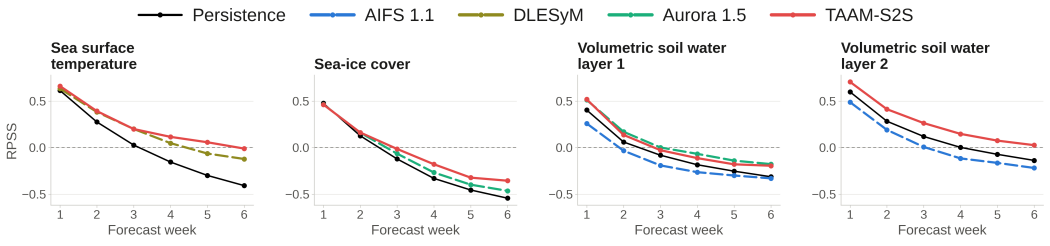

Figure 7: RPSS (↑) for slow variables achieved by the 12-member lagged ensembles.

Figure 8 shows the mean volumetric soil water layer 1-2 predictions in both regions. TAAM-S2S stays closer to ERA5 than the other models. This is especially visible for the Sichuan drought, where it keeps the anomaly longer. However, all models underestimate the drought intensity and revert towards climatology.

Table 3: Aggregated RMSE skill score (↑) with respect to climatology for the volumetric soil water layers 1-2 variables, for the 2022 Sichuan and 2023 Amazonia droughts.

	Sichuan 2022		Amazonia 2023	
Model	SWVL1	SWVL2	SWVL1	SWVL2
TAAM-S2S	**+0.162**	**+0.287**	**+0.274**	**+0.374**
AIFS v1.0/1.1	+0.009	-0.044	+0.194	+0.249
Aurora v1.5	+0.120	-	+0.198	-

To quantify the quality of the ensemble mean, we compute the latitude-weighted RMSE over each region at each lead time and take the root mean square over lead times (aggregated RMSE). The skill score against climatology is then $1 - AggRMSE_{pred}/AggRMSE_{climatology}$. As shown in Table 3, TAAM-S2S outperforms the competitors on both soil water layers during both droughts.

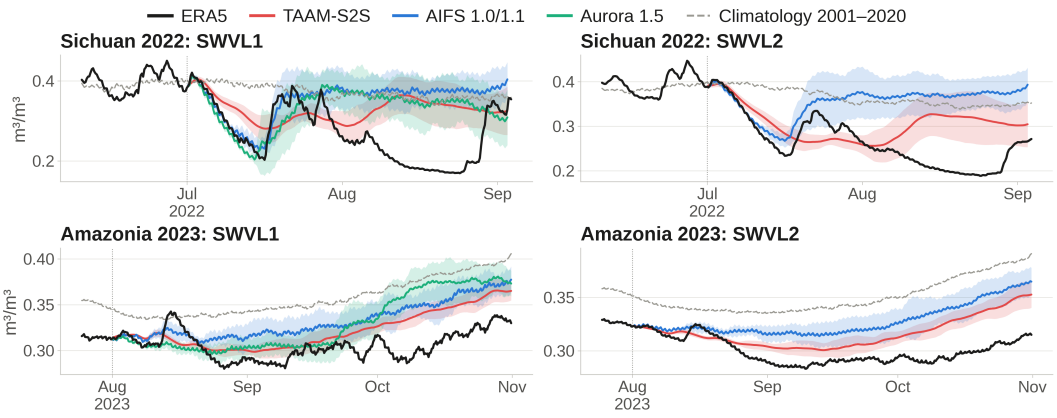

Figure 8: Predicted mean volumetric soil water layers 1-2 during the drought events in the Yangtze river basin in 2022 and Amazonia in 2023. The shaded areas represent ensemble standard deviations.

6 Conclusions and Limitations

S2S forecasting is a multi-scale learning problem where fast variables are well resolved as a single autoregressive step, but the evolution of slow variables is dominated by noise. We address this with two contributions. The Temporal Attention Aggregator (TAA) is a new temporal encoding mechanism that learns, for each variable and lead time, which past timesteps to attend to. Mixed-LT is a training and inference scheme that evolves slow and fast variables at separate paces. Together, they limit the issues arising from error accumulation and differing signal-to-noise ratio. Although neither is specific to a given model, we applied them to ESFM to obtain TAAM-S2S, which improves slow variable skill over the ESFM baselines while matching or improving fast variable skill. TAAM-S2S also outperforms AIFS v1.1, Aurora v1.5, and DLESyM in probabilistic scores for most slow variables, and better captures soil moisture anomalies during the 2022 Sichuan and 2023 Amazonia droughts. This comes at a fraction of their training cost, thanks to fewer atmospheric levels, a smaller model, and a coarser resolution that is sufficient for S2S timescales.

Yet several research directions remain open. While S2S forecasting often puts the emphasis on probabilistic predictions, TAAM-S2S is not intrinsically probabilistic and we compare either deterministic trajectories or ensembles created with lagged initialization. Capturing the intensity of rare events also remains a challenge, as all models underestimate the severity of the droughts studied in Section 5.2. Finally, we categorise variables into two classes (fast/slow), whereas in an ideal solution each feature should have its own lead time. Addressing this point is a promising direction towards a truly dynamics-aware model that evolves each feature at its own optimal pace.

AI USE STATEMENT

In this work, we used generative AI tools for assistance in writing code implementing part of the methods, and rarely for providing feedback on experiments and proposing interpretations of the results. We have not used generative AI tools for proposing or refining hypotheses, helping develop theoretical models or conceptual frameworks, designing research methodology or experiments, assisting with translation, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. The tasks of generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for spell-checking the article. We have reviewed all AI-assisted work. LLM-generated code was verified. The proposed results interpretations were always taken with the appropriate distance and lack of trust. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code used to obtain the presented results will be published together with the camera-ready version of the article. This includes scripts for data processing, model training and inference, calculating metrics, and creating plots. Data and models we compare with in this work are publicly available. Therefore, given enough compute, all of the results are reproducible.

REFERENCES

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep Learning for Day Forecasts from Sparse Observations. *arXiv preprint arXiv:2306.06079*, 2023.

Zied Ben Bouallègue, Mariana C. A. Clare, Linus Magnusson, Estibaliz Gascón, Michael Maier-Gerber, Martin Janoušek, Mark Rodwell, Florian Pinault, Jesper S. Dramsch, Simon T. K. Lang, Baudouin Raoult, Florence Rabier, Matthieu Chevallier, Irina Sandu, Peter Dueben, Matthew Chantry, and Florian Pappenberger. The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025.

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D Collins, Michael S Pritchard, and Alexander Keller. FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv preprint arXiv:2507.12144*, 2025.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. FuXi: A Cascade Machine Learning Forecasting System for 15-Day Global Weather Forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

Lei Chen, Xiaohui Zhong, Hao Li, Jie Wu, Bo Lu, Deliang Chen, Shang-Ping Xie, Libo Wu, Qingchen Chao, Chensen Lin, et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 15(1), 2024.

Nathaniel Cresswell-Clay, Bowen Liu, Dale R Durran, Zihui Liu, Zachary I Espinosa, Raul A Moreno, and Matthias Karlbauer. A deep learning earth system model for efficient simulation of the observed climate. *AGU Advances*, 6(4), 2025.

Cas Decancq, Thomas Mortier, Jessica Keune, and Diego G. Miralles. Weather Emulators at the Frontier of Heat Extremes Predictability. *arXiv preprint arXiv:2607.28220*, 2026.

Antoine Delaunay and Hannah M. Christensen. Interpretable Deep Learning for Probabilistic MJO Prediction. *Geophysical Research Letters*, 49(16):e2022GL098566, 2022.

James PC Duncan, Elynn Wu, Surya Dheeshjith, Adam Subel, Troy Arcomano, Spencer K Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W Andre Perkins, et al. SamudrACE: Fast and accurate coupled climate modeling with 3D ocean and atmosphere emulators. *Geophysical Research Letters*, 53(11), 2026.

Tom Dunstan, Oliver Strickson, Thusal Bennett, Jack Bowyer, Matthew Burnand, James Chappell, Alejandro Coca-Castro, Kirstine Ida Dale, Eric G Daub, Noushin Eftekhari, et al. FastNet: Improving the Physical Consistency of Machine Learning Weather Prediction Models through Loss Function Design. *Artificial Intelligence for the Earth Systems*, 5(3), 2026.

Jhan-Carlo Espinoza, Juan Carlos Jimenez, José Antonio Marengo, Jochen Schongart, Josyane Ronchail, Waldo Lavado-Casimiro, and João Vitor M Ribeiro. The new record of drought and warmth in the Amazon in 2023 related to regional and global climatic features. *Scientific Reports*, 14(1): 8107, 2024.

Hannah Guan, Soukayna Mouatadid, Paulo Orenstein, Judah Cohen, Haiyu Dong, Genevieve Elaine Flaspohler, Alex Xijie Lu, and Lester Mackey. Bridging the "predictability Desert"': A Probabilistic Bias Correction Framework for AI and Dynamical Subseasonal Forecasts. In *4th ICLR Workshop on Machine Learning for Remote Sensing (Main Track)*, 2026.

Sijie He, Xinyan Li, Timothy DelSole, Pradeep Ravikumar, and Arindam Banerjee. Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 169–177, 2021.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Nina Horat and Sebastian Lerch. Deep Learning for Postprocessing Global Probabilistic Forecasts on Subseasonal Time Scales. *Monthly Weather Review*, 152(3):667–687, 2024.

Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. Improving subseasonal forecasting in the western US with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2325–2335, 2019.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Hee-Seung Kim, Shanglin Zhou, Adam Bienkowski, and Krishna R. Pattipati. Subseasonal to Seasonal (S2S) Prediction Algorithms Using Hybrid Machine Learning Techniques. *Artificial Intelligence for the Earth Systems*, 4(3):e230108, 2025.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. AIFS–ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.

Guowen Li, Xintong Liu, Yang Liu, Mengxuan Chen, Shilei Cao, Xuehe Wang, Juepeng Zheng, Jinxiao Zhang, Haoyuan Liang, Lixian Zhang, et al. TianQuan-S2S: A Subseasonal-to-Seasonal Global Weather Model via Incorporating Climatology State. In *International Conference on Learning Representations*, volume 2026, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Olga Loegel, Joshua Talib, Frederic Vitart, Jörn Hoffmann, and Matthew Chantry. The AI Weather Quest: An International Competition for Sub-Seasonal Forecasting with AI. *Machine Learning: Earth*, 1(1):010701, 2025.

Riyu Lu, Ke Xu, Ruidan Chen, Wei Chen, Fang Li, and Chenyu Lv. Heat waves in summer 2022 and increasing concern regarding heat waves in general. *Atmospheric and Oceanic Science Letters*, 16(1):100290, 2023.

Tengfei Lyu, Weijia Zhang, and Hao Liu. Physics-Informed Teleconnection-Aware Transformer for Global Subseasonal-to-Seasonal Forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1054–1065. ACM, 2026.

Gabriel Moldovan, Ewan Pinnington, Ana Prieto Nemesio, Simon Lang, Zied Ben Bouallègue, Jesper Dramsch, Mihai Alexe, Mario Santa Cruz, Sara Hahner, Harrison Cook, et al. AIFS Single 1.1. 0: an update to ECMWF's machine-learned weather forecast model AIFS. *Geoscientific Model Development*, 19(10):4703–4724, 2026.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A Foundation Model for Weather and Climate. *arXiv preprint arXiv:2301.10343*, 2023.

Firat Ozdemir, Yun Cheng, Salman Mohebi, Fanny Lehmann, Simon Adamov, Zhenyi Zhang, Leonardo Trentini, Dana Grund, Oliver Fuhrer, Torsten Hoefler, et al. Earth System Foundation Model (ESFM): A unified framework for heterogeneous data integration and forecasting. *arXiv preprint arXiv:2605.00850*, 2026.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Vateanui Sansine, Takeshi Izumo, Marania Hopuare, Damien Specq, and Sophie Martinoni-La Pierre. Fine-Tuning a Global Weather Model for Superior Subseasonal Forecasting. 2025.

Michael Scheuerer, Matthew B. Switanek, Rochelle P. Worsnop, and Thomas M. Hamill. Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. *Monthly Weather Review*, 148(8):3489–3506, 2020.

Jakob Schloer, Steffen Tietsche, Christopher D. Roberts, Lorenzo Zampieri, Simon Lang, Gert Mertes, Gareth Jones, Matthew Chantry, and Frederic Vitart. AIFS-SUBS: Extending Data-Driven Forecasting to Sub-Seasonal Timescales. *arXiv preprint arXiv:2607.05100*, 2026.

Daisuke Tokuda and Paul A Dirmeyer. On the Comparison of Multiple Types of Soil Moisture Memory. *Journal of Hydrometeorology*, 27(5):635–650, 2026.

F. Vitart, C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H.-S. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A. W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D.-J. Won, H. Xiao, R. Zaripov, and L. Zhang. The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, 98(1):163–173, 2017.

Jonathan A. Weyn, Dale R. Durran, Rich Caruana, and Nathaniel Cresswell-Clay. Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models. *Journal of Advances in Modeling Earth Systems*, 13(7), 2021.

Jonathan A Weyn, Zekun Ni, Amit Misra, Will Fein, Haiyu Dong, Wessel P Bruinsma, Matt Corey, Kit Thambiratnam, Kevin White, Kenji Takeda, and Hongyu Sun. Aurora 1.5: Fine-Tuning a Foundation Model for Medium-Range Ensemble Weather Prediction. Technical report, Microsoft Publications, 2026.

Christopher J White, Henrik Carlsen, Andrew W Robertson, Richard JT Klein, Jeffrey K Lazo, Arun Kumar, Frederic Vitart, Erin Coughlan de Perez, Andrea J Ray, Virginia Murray, et al. Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological applications*, 24(3): 315–325, 2017.

Ziyu Zhou, Tian Zhou, Shiyu Wang, James Kwok, and Yuxuan Liang. SOON: Symmetric Orthogonal Operator Network for Global Subseasonal-to-Seasonal Climate Forecasting. *arXiv preprint arXiv:2602.15040*, 2026.

A TRAINING PROCEDURE

Each model is trained using one NVIDIA GH200 compute node containing four 96 GB GPUs. For the training of the ESFM baselines and mixed-LT models we set the batch size to 4 and optimize the model with AdamW and a maximal learning rate of 5e-4. We use a cosine learning rate scheduler starting from a 1000-step warmup period with a linear increase. Unless stated otherwise, we use the latitude- and variable-weighted MAE between predictions $\hat{\boldsymbol{X}}$ and ground truth $\boldsymbol{X}$ as the loss:

$$\mathcal{L}_{MAE}(\hat{\boldsymbol{X}}, \boldsymbol{X}) = \frac{1}{|V|} \sum_{v \in V} w_v \cdot \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} a_i \left| \hat{\boldsymbol{X}}_{i,j}^{(v)} - \boldsymbol{X}_{i,j}^{(v)} \right|$$

where V is the set of variables, w_v is the variable weight, set by default to $\frac{1}{\sigma_v}$, H and W are the grid height and width, and a_i is the latitude weight defined by:

$$a_i = \frac{\cos \phi_i}{\frac{1}{H} \sum_{i'=1}^{H} \cos \phi_{i'}}$$

with ϕ_i being the latitude of row i.

The **baseline** ESFM models are trained for 100 epochs, where one epoch contains four timesteps (00:00, 06:00, 12:00, 18:00 UTC) for each day between 1979 and 2018. The **mixed-LT** models we train for just 10 epochs, as they are initialised from already trained baselines. We also observe some overfitting for a few variables, such as sea surface variables and volumetric soil water layers 2-4, in the longer trainings.

We run **block rollout finetuning** for 5000 steps. Here, the maximal linear learning rate was set to 5e-5 and the linear warmup had 200 steps. Furthermore, in this setting we add a modified spherical harmonic (MSH) component, which changes the loss to:

$$\mathcal{L}_{rollout}(\hat{\boldsymbol{X}}, \boldsymbol{X}) = \mathcal{L}_{MAE}(\hat{\boldsymbol{X}}, \boldsymbol{X}) + w_{MSH} \mathcal{L}_{MSH}(\hat{\boldsymbol{X}}, \boldsymbol{X})$$

For the set of fast variables F we have $\mathcal{L}_{MSH}(\hat{\boldsymbol{X}}, \boldsymbol{X}) = \frac{1}{|F|} \sum_{v \in F} MSH(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)})$, and:

$$MSH(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)}) = \sum_k \gamma_k \Bigg[\left(\sqrt{PSD_k(\hat{\boldsymbol{X}}^{(v)})} - \sqrt{PSD_k(\boldsymbol{X}^{(v)})} \right)^2 + 2 \max\Big(PSD_k(\hat{\boldsymbol{X}}^{(v)}), PSD_k(\boldsymbol{X}^{(v)}) \Big) \Big(1 - Coh_k(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)}) \Big) \Bigg],$$

where:

$$\gamma_k = \max\left(\left(\frac{k}{k_0} \right)^{\sqrt{3}}, 1 \right)$$

$$PSD_k(\boldsymbol{X}^{(v)}) = \sum_l \left| \alpha_{X^{(v)}}(k,l) \right|^2,$$

$$Coh_k(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)}) = \frac{\sum_l \mathrm{Re}\Big(\alpha_{X^{(v)}}(k,l) \cdot \overline{\alpha_{X^{(v)}}(k,l)} \Big)}{\sqrt{PSD_k(\hat{\boldsymbol{X}}^{(v)})\, PSD_k(\boldsymbol{X}^{(v)})}}.$$

In the formulas above, $\alpha_{X^{(v)}}(k,l)$ is the spectral coefficient corresponding to the total wavenumber k and zonal wavenumber l. We choose $k_0 = 20$ (wavelength $\sim$2000 km) to put more emphasis on small-scale features. The γ_k values are plotted in Figure 9. w_{MSH} we set to obtain a selected MAE:MSH loss contribution ratio. To calculate it, we compare the total MAE and MSH on 40 random samples from the training set. For more details about the MSH loss, see the work by Dunstan et al. (2026).

Including setup, exploratory experiments, ablation studies, and the described results across the whole duration of the project, we estimate that $\sim$ 8000 GPU hours were used. A full 100-epoch baseline training took approximately 30h on one NVIDIA GH200 node, and a 10-epoch mixed-LT training roughly 5h in the same setup. Rollout finetuning, despite running for only 5000 steps with a batch size of one, required up to 18h using four GPUs. Data processing and inference costs were negligible in comparison to the training cost.

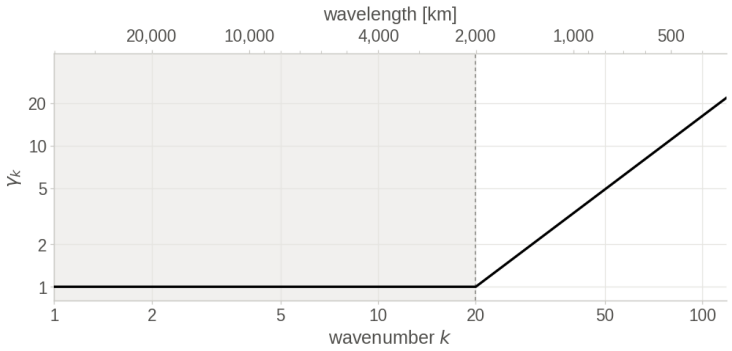

Figure 9: γ_k values corresponding to $k_0 = 20$. This choice ensures that the emphasis is put on small scale features. Note that both axes on the plot are logarithmic.

B DATA

ERA5 variables used for the model training are listed in Table 4. All of our models support the same set of variables, including eleven surface variables, five atmospheric ones (on three levels: 1000 hPa, 850 hPa and 500 hPa) and three static variables. The model omits the static variables in the output. Six of the surface variables we classify as slowly-evolving. We distinguish only two classes of variable dynamics, fast and slow, even though there is still significant variability between their paces, as, for example, volumetric soil water layer 4 changes noticeably slower than layer 1.

We predict precipitation as a logarithm for better model performance, and always in 6h windows, despite the different lead times, for comparability. Land and sea masks for slow variables are derived from the land sea mask static variable. Sea surface temperature and sea ice are provided as daily values in the ERA5 dataset. For our training purposes, we assume that those correspond to the feature maps around 12:00 UTC, and interpolate them linearly at the timesteps in between to obtain a 6h cadence.

Table 4: Variables used during model training and evaluation.

	Variable	Abbr.	Units	Dynamic	Mask
Surface	2m temperature	2T	K	fast	–
	10m U-component of wind	10U	$\mathrm{m\,s^{-1}}$	fast	–
	10m V-component of wind	10V	$\mathrm{m\,s^{-1}}$	fast	–
	Mean sea-level pressure	MSL	Pa	fast	–
	Sea surface temperature	SST	K	slow	land
	Sea-ice cover	CI	–	slow	land
	Volumetric soil water layer 1 (0-7 cm)	SWVL1	$\mathrm{m^3\,m^{-3}}$	slow	sea
	Volumetric soil water layer 2 (7-28 cm)	SWVL2	$\mathrm{m^3\,m^{-3}}$	slow	sea
	Volumetric soil water layer 3 (28-100 cm)	SWVL3	$\mathrm{m^3\,m^{-3}}$	slow	sea
	Volumetric soil water layer 4 (100-289 cm)	SWVL4	$\mathrm{m^3\,m^{-3}}$	slow	sea
	6-hourly total precipitation (log)	TP6h	–	fast	–
Atmosphere	Geopotential	Z	$\mathrm{m^2\,s^{-2}}$	fast	–
	U-component of wind	U	$\mathrm{m\,s^{-1}}$	fast	–
	V-component of wind	V	$\mathrm{m\,s^{-1}}$	fast	–
	Temperature	T	K	fast	–
	Specific humidity	Q	$\mathrm{kg\,kg^{-1}}$	fast	–
Static	Land-sea mask	LSM	–	const.	–
	Surface geopotential	Z	$\mathrm{m^2\,s^{-2}}$	const.	–
	Soil type	SLT	–	const.	–

C METRICS

Throughout the work, we focus on three evaluation metrics.

Root mean square error (RMSE) is used to capture the pixel-wise accuracy of the model, looking at each predicted value separately. Since grid points closer to the equator correspond to larger areas, we use latitude weighting:

$$\mathrm{RMSE}(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)}) = \sqrt{\frac{\sum_{i=1}^{H}\sum_{j=1}^{W} a_i \left(\hat{\boldsymbol{X}}_{i,j}^{(v)} - \boldsymbol{X}_{i,j}^{(v)}\right)^2}{\sum_{i=1}^{H}\sum_{j=1}^{W} a_i}}$$

RMSE is closely related to the MAE loss we optimised during training. The downside of this metric is that it allows the model to produce blurred predictions in order to optimize it. RMSE alone therefore cannot distinguish a genuinely skilful forecast from an overly smooth one, which is why it is reported alongside the two metrics below.

Anomaly correlation coefficient (ACC) measures how well the forecast reproduces deviations from the seasonal norm. Anomalies are taken with respect to a 20-year (2001–2020) ERA5 climatology resolved for each day of year and hour ($\boldsymbol{C}$):

$$\mathrm{ACC}(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)}) = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} a_i \hat{\boldsymbol{A}}_{i,j}^{(v)} \boldsymbol{A}_{i,j}^{(v)}}{\sqrt{\sum_{i=1}^{H}\sum_{j=1}^{W} a_i \left[\hat{\boldsymbol{A}}_{i,j}^{(v)}\right]^2 \cdot \sum_{i=1}^{H}\sum_{j=1}^{W} a_i \left[\boldsymbol{A}_{i,j}^{(v)}\right]^2}}, \qquad \boldsymbol{A} = \boldsymbol{X} - \boldsymbol{C}.$$

Because it is normalised, ACC is dimensionless and comparable across variables and lead times. Moreover, it isolates the part of the signal that is more informative, as it measures performance with the predictable seasonal component removed. A climatological forecast has zero anomaly by definition, and therefore zero ACC.

Logarithmic spectral error (LSE) represents the root mean square error between the log-transformed power spectral densities of the predictions and of the ground truth. To calculate it, for each feature map the field's spatial mean is removed, then masked cells are set to zero, and a Hann window is applied in latitude to account for the lack of periodicity at the north-south edges. We then apply a 2D FFT and square the resulting coefficients, which gives the power of every pair of zonal and meridional wavenumbers (i, j). Those powers are radially averaged over the total wavenumber $k = \sqrt{i^2 + j^2}$, discarding orientation, which yields the power spectral density P_k as a function of spatial scale.

$$\mathrm{LSE}(\hat{\boldsymbol{X}}^{(v)}, \boldsymbol{X}^{(v)}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log \mathrm{P}_k(\hat{\boldsymbol{X}}^{(v)}) - \log \mathrm{P}_k(\boldsymbol{X}^{(v)})\right)^2} \tag{3}$$

LSE measures how much structure the forecast has at each scale. Importantly for our use case, the logarithmic transformation gives us a metric that accounts for small-scale errors, even though they have significantly smaller magnitudes.

D TEMPORAL ATTENTION AGGREGATOR (TAA)

During mixed-LT training, when we replace the original ESFM patch embedding with TAA, we also randomly re-initialise the whole encoder. Inside TAA, both the relative time embedding Δ and the lead time embedding γ are obtained by encoding their respective values in hours using Fourier expansions followed by linear layers of size 256×256. The Fourier expansion covers wavelengths from 1/60h to 504h. In each experiment, we use four attention heads in TAA, each of dimension 64. The MLP that processes the embedding after adding the residual and applying layer normalisation has a hidden dimension of 1024 and uses the GELU activation function.

E TAA MODEL FOR S2S FORECASTING (TAAM-S2S)

Figures 10, 11 and 12 present the ACC, LSE and RMSE metrics for the TAAM-S2S model compared with the 6h and 48h baselines. TAAM-S2S outperforms or matches the baseline models on the vast majority of variables in terms of ACC and RMSE. LSE performance is variable-dependent. We also note a visible degradation in the metrics for the 6-hourly total precipitation variable. When optimizing the MSH component, the models produce spikes at individual pixels, and since we natively predict this variable in log space, we obtain unrealistic values after the inverse transformation. Consequently, the plotted values do not reflect the true performance of the model for this variable.

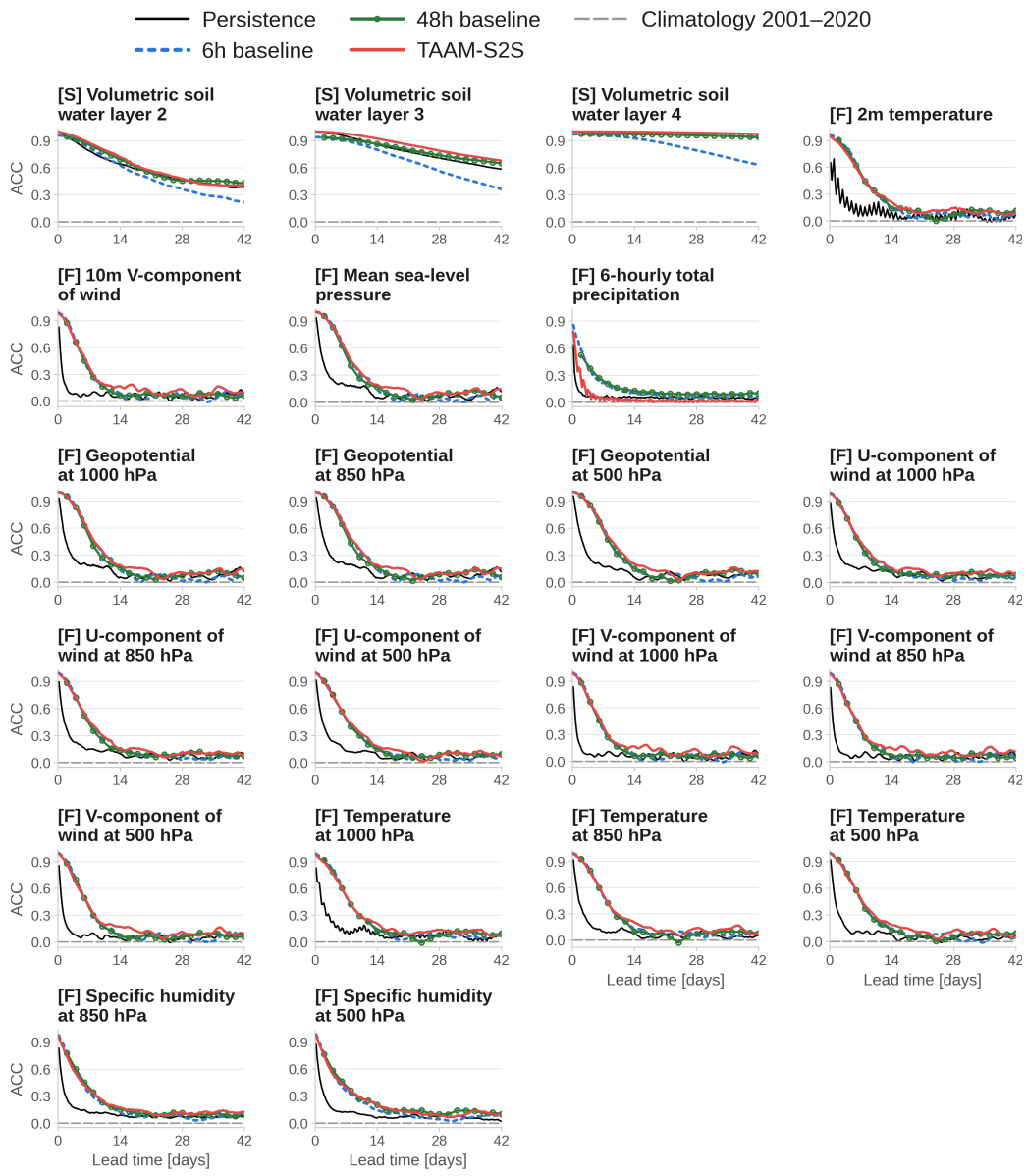

Figure 10: ACC (↑) as a function of lead time for selected slow and fast variables, scoring TAAM-S2S against ESFM-based baselines, persistence and climatology. [F] denotes a fast variable and [S] a slow variable.

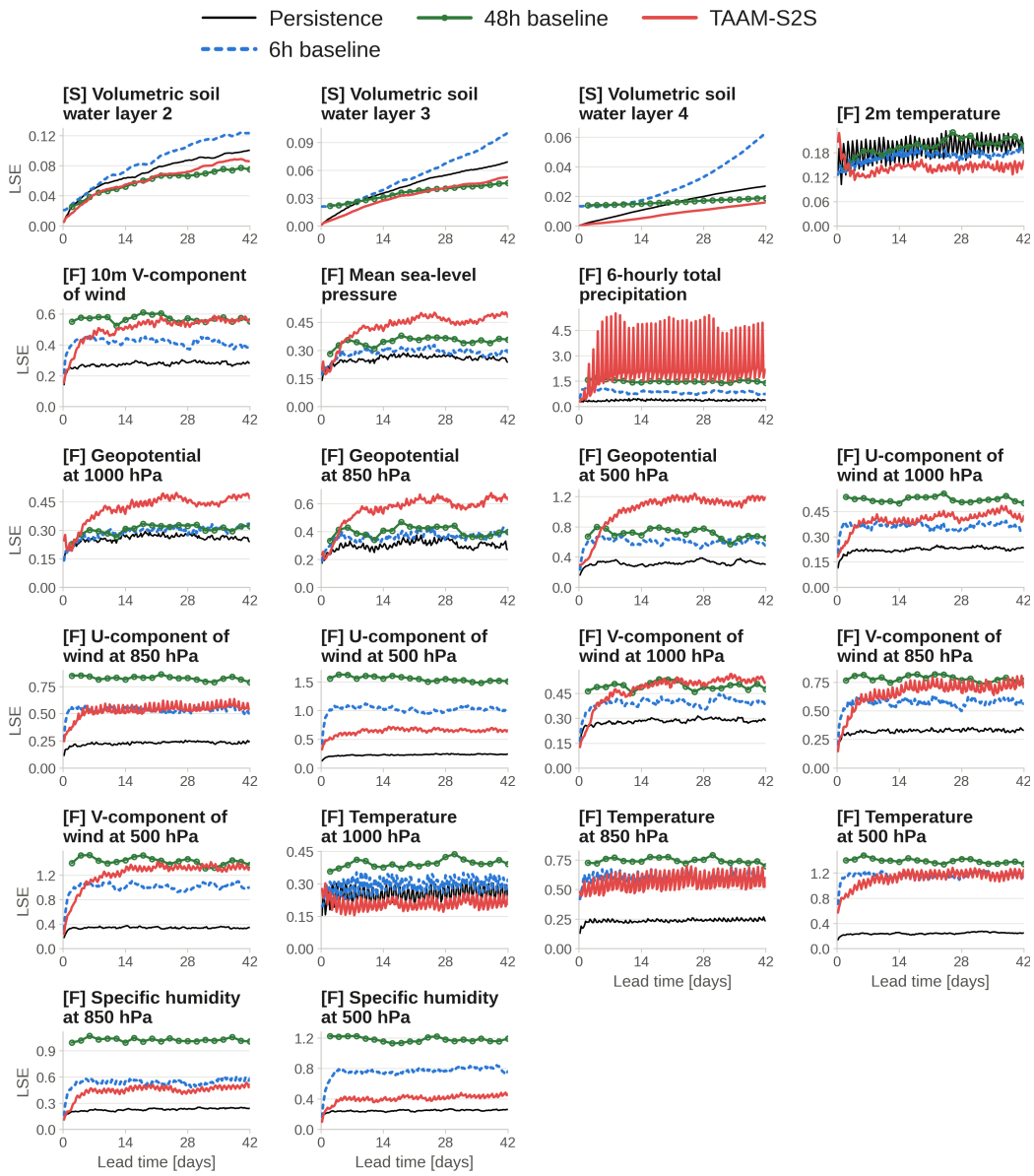

Figure 11: LSE (↓) as a function of lead time for selected slow and fast variables, scoring TAAM-S2S against ESFM-based baselines, persistence and climatology. [F] denotes a fast variable and [S] a slow variable.

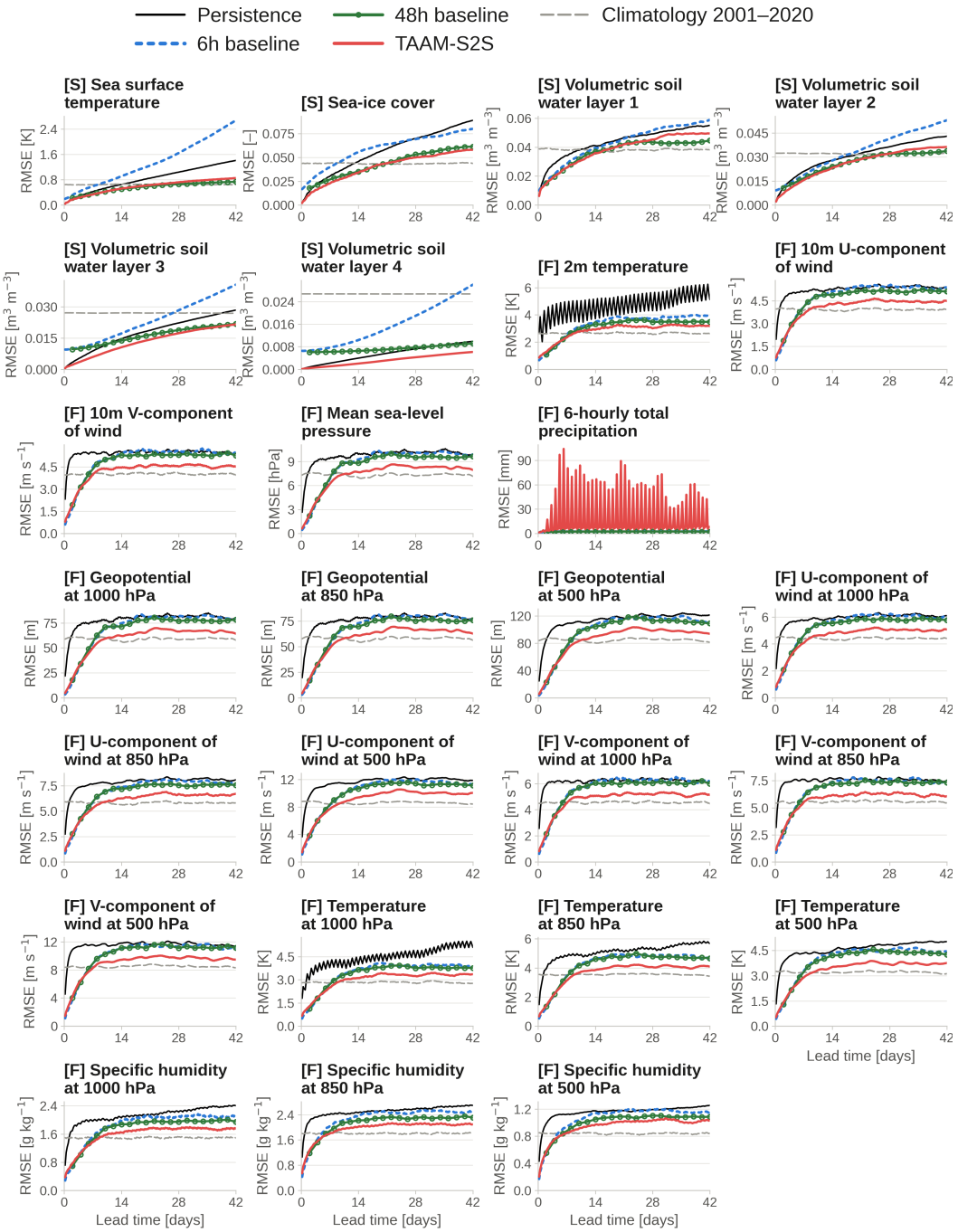

Figure 12: RMSE (↓) as a function of lead time for selected slow and fast variables, scoring TAAM-S2S against ESFM-based baselines, persistence and climatology. [F] denotes a fast variable and [S] a slow variable.

F MIXED LEAD TIME SELECTION

Tables 5 and 6 present the RMSE and LSE metrics for the mixed-LT models with different selections of $\delta_2 \in \{48\text{h}, 96\text{h}, 192\text{h}\}$. The 6/48h and 6/192h mixed-LT models show promising performance in terms of RMSE on slow variables. However, of these two, only the 6/48h model is competitive with the 6h baseline on fast variables at mid-range time horizons, in terms of both RMSE and LSE.

Table 5: Mean and standard deviation of RMSE (↓) for slow and fast variables calculated based on 12 S2S rollouts initialized in 2021. Best scores are highlighted in bold, second best are underlined. Values are presented in variables' native units listed in Table 4.

Model	Slow variables (weeks 1–6)					
	SST	CI	SWVL1	SWVL2	SWVL3	SWVL4
6 h baseline	1.33 ± 0.11	0.059 ± 0.008	0.043 ± 0.003	0.034 ± 0.002	0.023 ± 0.001	0.016 ± 0.001
48 h baseline	0.56 ± 0.03	0.044 ± 0.007	0.037 ± 0.003	0.027 ± 0.003	<u>0.016</u> ± 0.001	**0.007** ± 0.000
mixed-LT 6/48 h	<u>0.54</u> ± 0.04	0.046 ± 0.005	<u>0.036</u> ± 0.003	<u>0.026</u> ± 0.003	0.017 ± 0.001	0.010 ± 0.001
mixed-LT 6/96 h	0.60 ± 0.04	<u>0.042</u> ± 0.004	0.037 ± 0.002	0.027 ± 0.002	0.017 ± 0.001	0.010 ± 0.000
mixed-LT 6/192 h	**0.49** ± 0.02	**0.037** ± 0.004	**0.033** ± 0.002	**0.024** ± 0.002	**0.015** ± 0.001	<u>0.008</u> ± 0.000
Model	Fast variables (weeks 1–2)					
	2T	TP6h	T850	Z500	V1000	Q1000
6 h baseline	**2.34** ± 0.25	<u>2.08</u> ± 0.05	**3.11** ± 0.26	**66.97** ± 5.20	**4.38** ± 0.18	**1.38** ± 0.05
48 h baseline	<u>2.43</u> ± 0.29	2.10 ± 0.06	3.35 ± 0.30	74.72 ± 7.20	4.73 ± 0.26	1.44 ± 0.07
mixed-LT 6/48 h	2.81 ± 0.30	2.13 ± 0.07	<u>3.18</u> ± 0.25	<u>68.59</u> ± 5.14	<u>4.45</u> ± 0.20	<u>1.43</u> ± 0.05
mixed-LT 6/96 h	2.87 ± 0.33	2.08 ± 0.05	3.66 ± 0.32	81.04 ± 5.09	4.81 ± 0.17	1.57 ± 0.06
mixed-LT 6/192 h	2.81 ± 0.36	**2.08** ± 0.04	3.53 ± 0.31	76.64 ± 4.23	4.71 ± 0.12	1.53 ± 0.03

Table 6: Mean and standard deviation of LSE (↓) for slow and fast variables calculated based on 12 S2S rollouts initialized in 2021. Best scores are highlighted in bold, second best are underlined.

Model	Slow variables (weeks 1–6)					
	SST	CI	SWVL1	SWVL2	SWVL3	SWVL4
6 h baseline	0.062 ± 0.008	0.225 ± 0.073	0.106 ± 0.007	0.086 ± 0.007	0.054 ± 0.008	0.029 ± 0.003
48 h baseline	0.035 ± 0.007	0.196 ± 0.033	**0.080** ± 0.008	**0.059** ± 0.004	<u>0.036</u> ± 0.002	**0.016** ± 0.002
mixed-LT 6/48 h	<u>0.032</u> ± 0.006	0.084 ± 0.026	0.097 ± 0.010	0.068 ± 0.007	0.038 ± 0.004	0.027 ± 0.004
mixed-LT 6/96 h	0.042 ± 0.009	<u>0.080</u> ± 0.034	0.110 ± 0.010	0.078 ± 0.007	0.041 ± 0.004	0.017 ± 0.001
mixed-LT 6/192 h	**0.030** ± 0.003	**0.066** ± 0.027	<u>0.088</u> ± 0.013	<u>0.062</u> ± 0.007	**0.035** ± 0.004	<u>0.017</u> ± 0.001
Model	Fast variables (weeks 1–2)					
	2T	TP6h	T850	Z500	V1000	Q1000
6 h baseline	**0.159** ± 0.021	**0.924** ± 0.122	**0.594** ± 0.045	**0.602** ± 0.066	**0.373** ± 0.021	**0.221** ± 0.027
48 h baseline	<u>0.175</u> ± 0.033	1.550 ± 0.121	0.742 ± 0.049	0.724 ± 0.059	0.485 ± 0.031	0.317 ± 0.044
mixed-LT 6/48 h	0.192 ± 0.028	<u>1.252</u> ± 0.202	<u>0.710</u> ± 0.048	<u>0.656</u> ± 0.056	<u>0.453</u> ± 0.021	<u>0.293</u> ± 0.038
mixed-LT 6/96 h	0.216 ± 0.033	1.616 ± 0.171	0.840 ± 0.051	0.853 ± 0.067	0.622 ± 0.042	0.340 ± 0.041
mixed-LT 6/192 h	0.209 ± 0.033	1.500 ± 0.176	0.821 ± 0.061	0.773 ± 0.085	0.592 ± 0.045	0.333 ± 0.057

G INCREASING THE NUMBER OF LEAD TIMES

To verify the effectiveness of TAA and the mixed-LT approach on more than two lead times, we trained one model to predict 6h, 48h, 96h, 192h and 384h steps at the same time (mixed-LT 6/48/96/192/384h). It was initialised from the 6h baseline and trained for 20 epochs instead of 10 to account for the greater number of tasks the model had to learn. We evaluated this model using the same inference scheme as described in Section 3.3, where we choose only one δ_2. Note, however, that if we divided our variables into more than two classes (fast/slow), we could perform rollouts using more lead times.

Figures 13 and 14 compare the RMSE of the mixed-LT 6/48/96/192/384h model with that of mixed-LT 6/48h and mixed-LT 6/192h, respectively, using the same δ_2. The model with an increased number of learned lead times usually scores slightly worse on slow variables, but better on fast variables. However, the improved results on RMSE do not come from greater skill, but from mildly blurred or unphysical predictions. It is reflected in increased LSE visible in Figures 15 and 16.

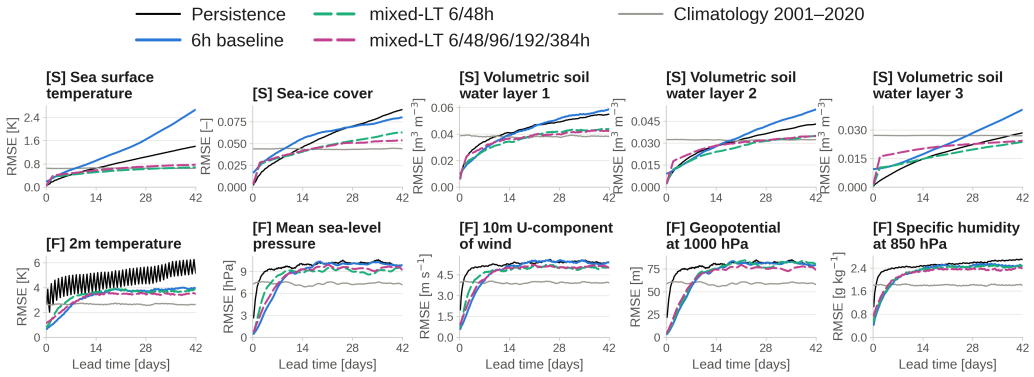

Figure 13: RMSE (↓) as a function of lead time comparing mixed-LT 6/48h and mixed-LT 6/48/96/192/384h models' rollouts created with $\delta_2 = 48$h.

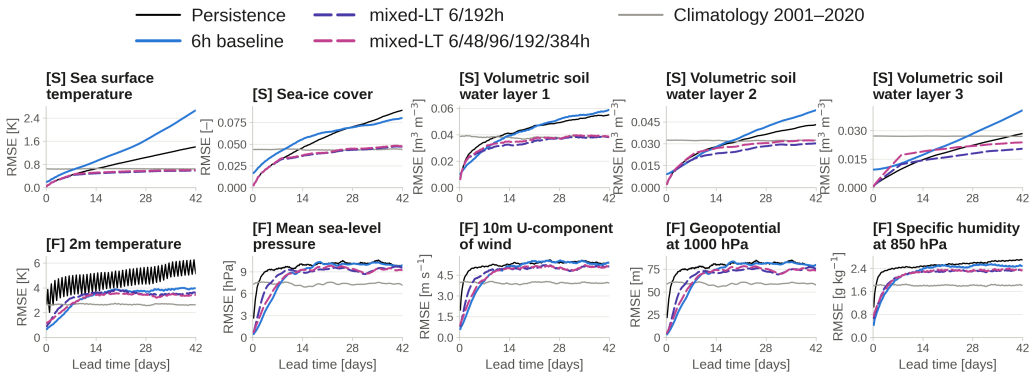

Figure 14: RMSE (↓) as a function of lead time comparing mixed-LT 6/192h and mixed-LT 6/48/96/192/384h models' rollouts created with $\delta_2 = 192$h.

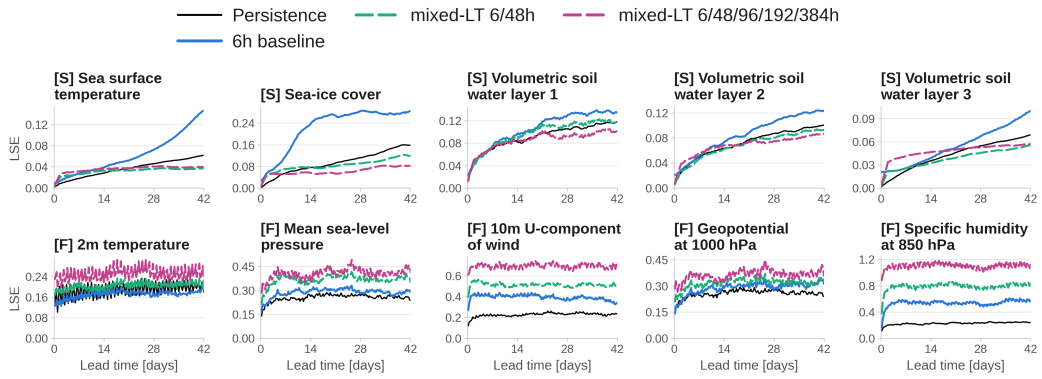

Figure 15: LSE (↓) as a function of lead time comparing mixed-LT 6/48h and mixed-LT 6/48/96/192/384h models' rollouts created with $\delta_2 = 48$h.

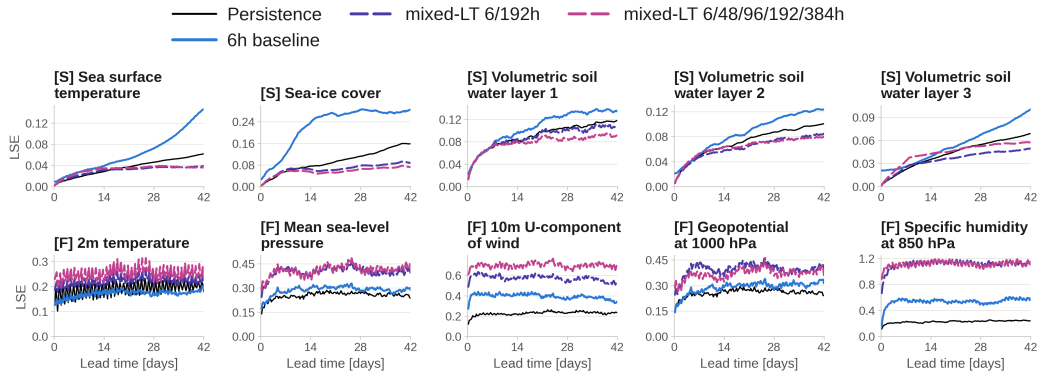

Figure 16: LSE (↓) as a function of lead time comparing mixed-LT 6/192h and mixed-LT 6/48/96/192/384h models' rollouts created with $\delta_2 = 192$h.

H RESIDUAL PREDICTIONS

When training mixed-LT models with slow-variable residuals as the target, we initialise from the 6h and 48h baselines trained on the same tasks. For these models, we recomputed the mean and standard deviation statistics of the predicted values used for normalization, separately for each lead time. Table 7 compares the ACC scores on slow variables obtained using absolute and residual models. Using residuals substantially improves performance for the majority of slow variables in all tested models.

Figure 17 shows an example of sea-ice cover prediction made by two mixed-LT models trained with lead times $\delta_1 = 6\text{h}$ and $\delta_2 = 48\text{h}$. The models differ in the baseline model chosen for initialization. The model initialised from the 6h baseline was not able to learn sea-ice dynamics and instead produced persistence forecasts. This issue was overcome by initialising from the 48h baseline, which had already learned to model sea-ice cover during the first training phase. This behaviour was observed for all tested $\delta_2 \in \{48\text{h}, 96\text{h}, 192\text{h}\}$.

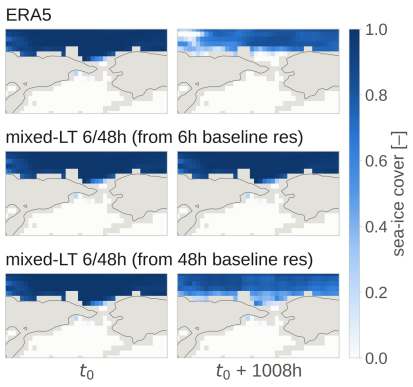

Figure 17: ERA5 and predicted sea-ice cover over the Bering Strait. t_0 and $t_0 + 1008\text{h}$ forecasts are identical for the mixed-LT model initialized from the 6h residual baseline.

We switched from an absolute to a residual target only for the slow variables. We also conducted experiments applying the residual strategy to the fast variables. In this case, however, we observed patch artifacts in longer forecasts (Figure 18), caused by compounding systematic biases in the residual predictions.

Table 7: Mean and standard deviation of ACC (↑) for the baselines and mixed-LT 6/48h model predicting absolute feature maps and their equivalents optimised to predict residuals for slow variables.

Model	sea surface temperature		sea-ice cover		volumetric soil water layer 1	
	Weeks 1–2	Weeks 3–6	Weeks 1–2	Weeks 3–6	Weeks 1–2	Weeks 3–6
6 h baseline (absolute)	0.65 ± 0.06	0.21 ± 0.08	0.60 ± 0.06	0.13 ± 0.12	0.70 ± 0.06	0.27 ± 0.08
6 h baseline (residual)	$\underline{0.83} \pm 0.04$	0.41 ± 0.10	0.76 ± 0.05	0.19 ± 0.18	$\underline{0.72} \pm 0.06$	0.26 ± 0.07
48 h baseline (absolute)	0.80 ± 0.04	0.36 ± 0.10	0.77 ± 0.05	0.33 ± 0.14	0.70 ± 0.05	0.33 ± 0.10
48 h baseline (residual)	0.83 ± 0.04	$\underline{0.47} \pm 0.07$	$\underline{0.84} \pm 0.03$	$\underline{0.38} \pm 0.09$	0.71 ± 0.07	$\underline{0.36} \pm 0.09$
mixed-LT 6/48 h (absolute)	0.76 ± 0.05	0.46 ± 0.10	0.70 ± 0.08	0.27 ± 0.11	0.72 ± 0.06	$\mathbf{0.37} \pm 0.09$
mixed-LT 6/48 h (residual)	$\mathbf{0.85} \pm 0.03$	$\mathbf{0.47} \pm 0.07$	$\mathbf{0.84} \pm 0.03$	$\mathbf{0.40} \pm 0.12$	$\mathbf{0.75} \pm 0.06$	0.33 ± 0.09

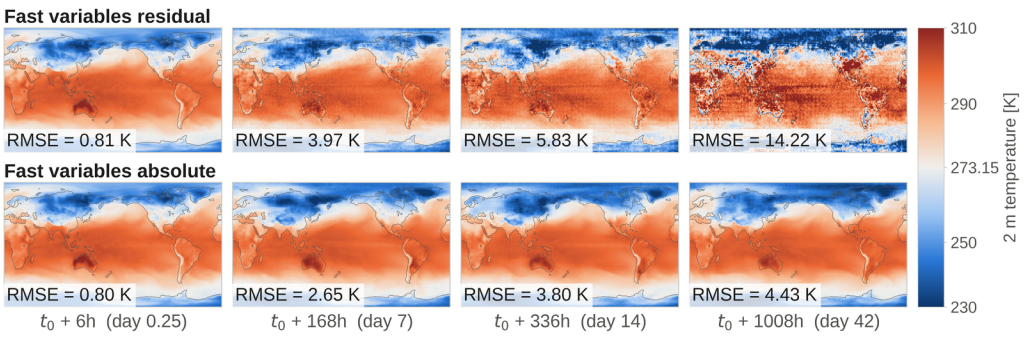

Figure 18: Comparison of 2-meter temperature rollouts from the mixed-LT 6/48h models. The top row shows forecasts from the model predicting residual maps across all variables, while the bottom row displays the corresponding rollout from the model predicting residuals solely for slow variables.

I 10-YEAR ROLLOUTS

This section presents a study of the models' stability over a 10-year rollout. Figure 19 shows the monthly global means of predicted sea surface temperature, sea-ice cover, 2m temperature and specific humidity at 850 hPa starting from January 2020. We compare the 48h baseline and mixed-LT 6/48h, including two variants of each model: one predicting absolute values of slow variables and one predicting residuals.

For sea surface temperature, we observe that the 48h baseline with residual slow variables performs significantly better than its absolute equivalent. In sea-ice cover prediction, only the mixed-LT 6/48h absolute model seems to be keeping the correct trend. Interestingly, for both of those variables we can see that the non-residual models are losing the seasonality, as can be observed in the much smaller oscillations.

The latter observation is even stronger for the selected fast variables. There, all of the rollouts stay within feasible values, but for 2m temperature and specific humidity at 850 hPa, the absolute models have significantly smaller amplitudes than the reference. We do not observe this for residual models.

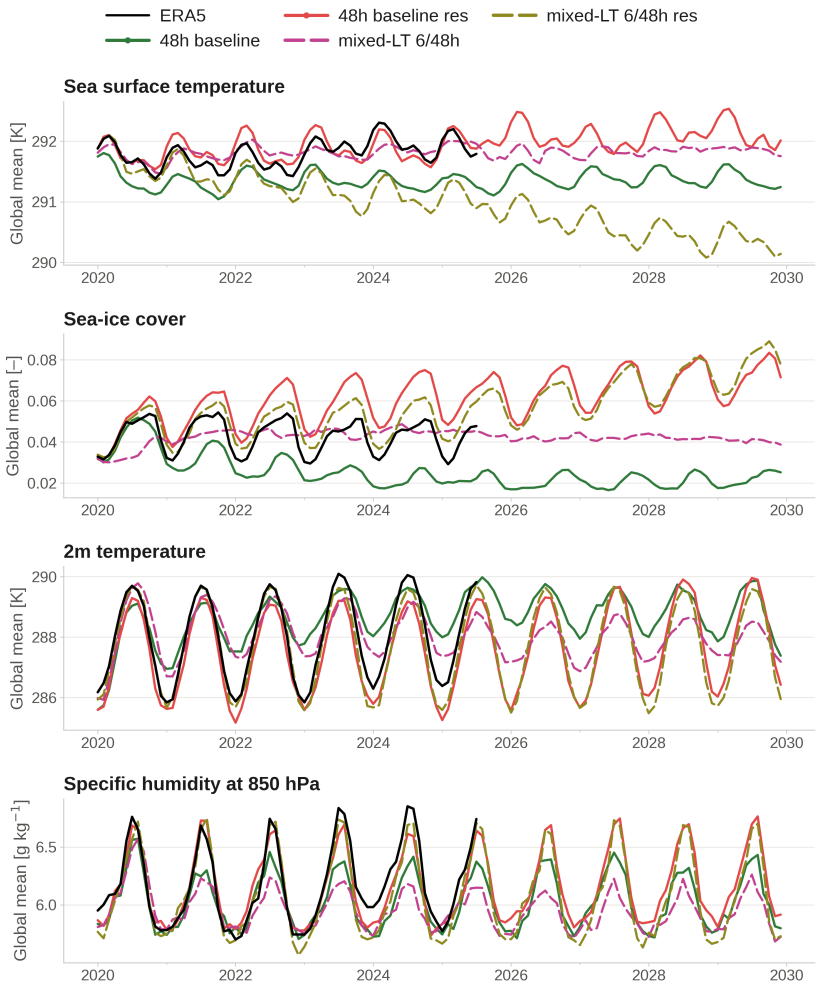

Figure 19: Monthly global means of predicted sea surface temperature, sea-ice cover, 2m temperature, and specific humidity at 850 hPa. They were derived from 10-year rollouts starting from January 2020. We include predictions of the 48h baseline and mixed-LT 6/48h, in variants predicting slow variables as absolute values and as residuals (denoted "res").

J BLOCK ROLLOUT FINETUNING

Figures 20 and 21 present the RMSE and LSE metrics after block rollout finetuning. We observe that training with a 9:1 MAE:MSH ratio results in a slightly better RMSE than TAAM-S2S (5:5). However, TAAM-S2S has significantly improved LSE compared to the 9:1 model. We also observe the same degradation in precipitation metrics as described in Appendix E.

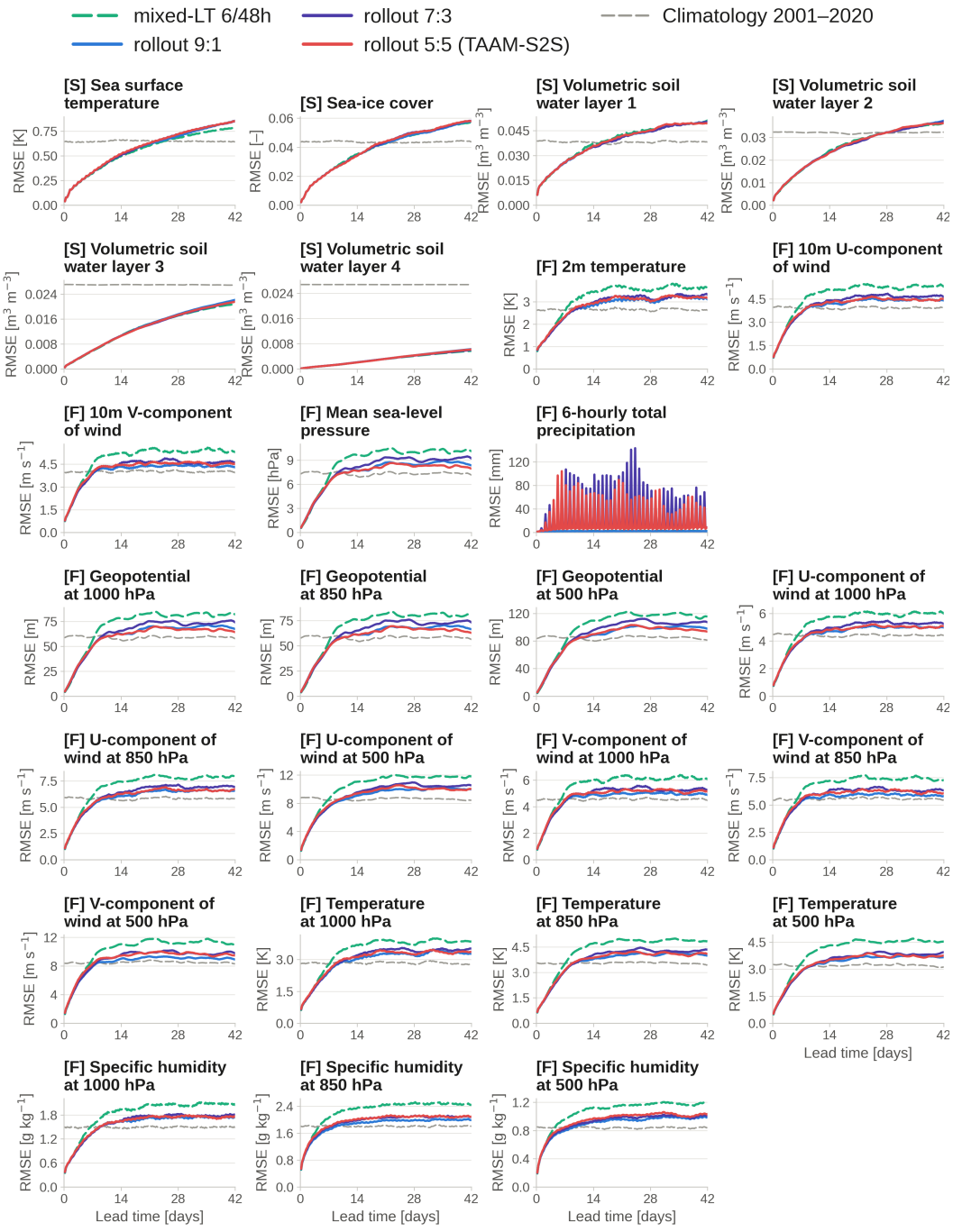

Figure 20: RMSE (↓) as a function of lead time for mixed-LT models after block rollout finetuning with different MAE:MSH loss ratio. [F] denotes a fast variable and [S] a slow variable.

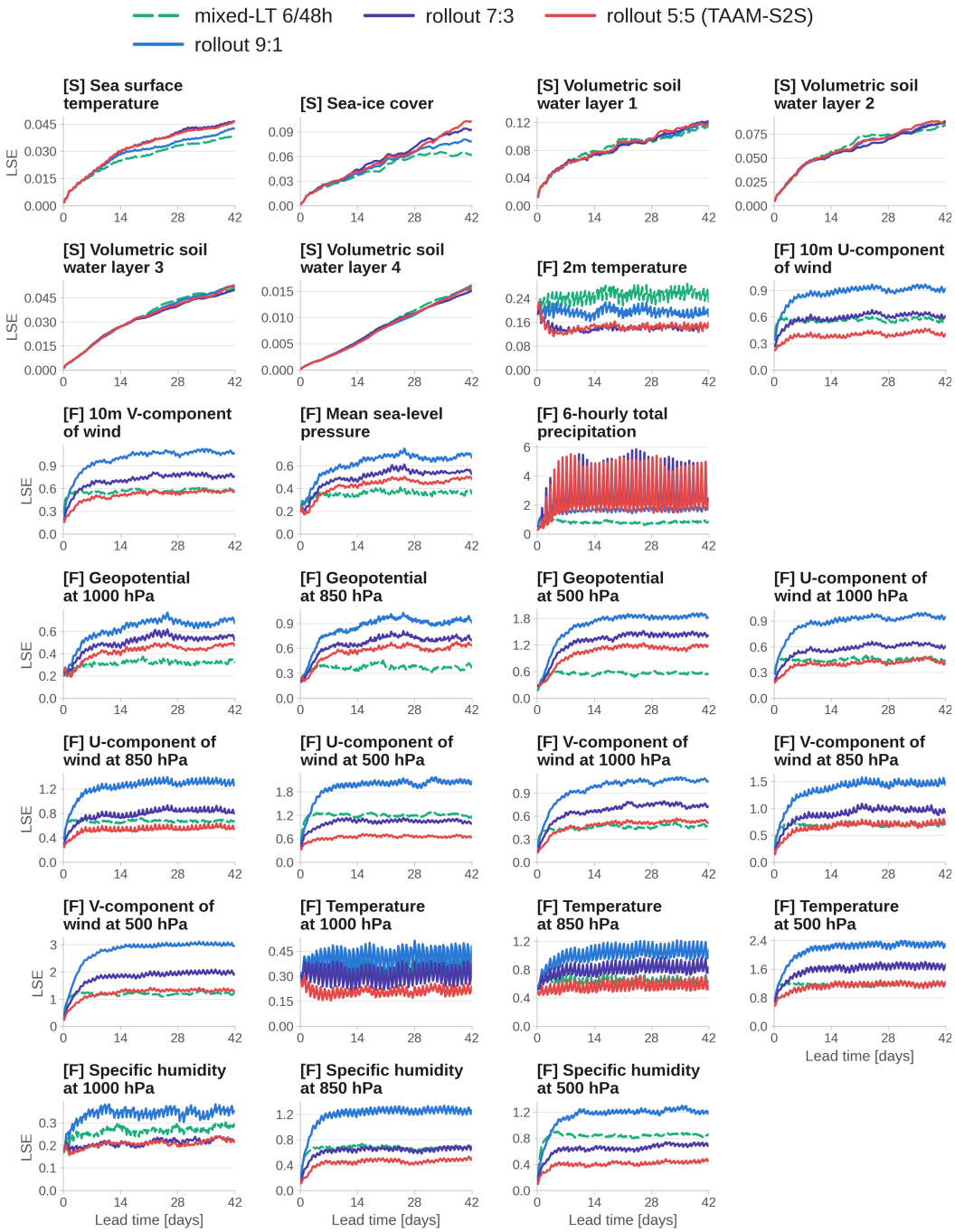

Figure 21: LSE (↓) as a function of lead time for mixed-LT models after block rollout finetuning with different MAE:MSH loss ratio. [F] denotes a fast variable and [S] a slow variable.

Figures 22 and 23 contain the power spectra of predicted variables after rollout finetuning, for 6h and 1008h lead times, respectively.

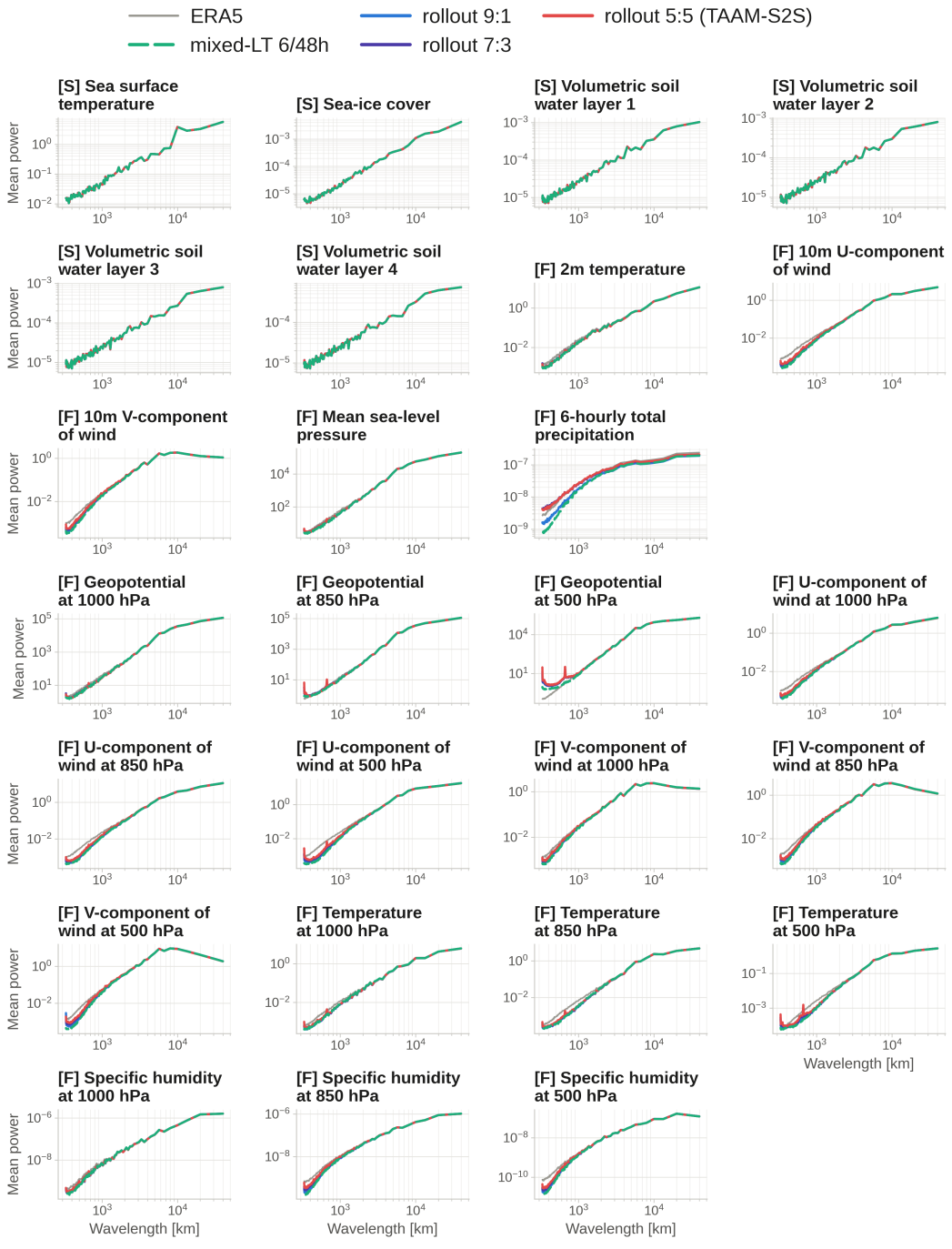

Figure 22: Power spectra of variables predicted at 6h lead time by the rollout-finetuned models and the mixed-LT 6/48h residual model.

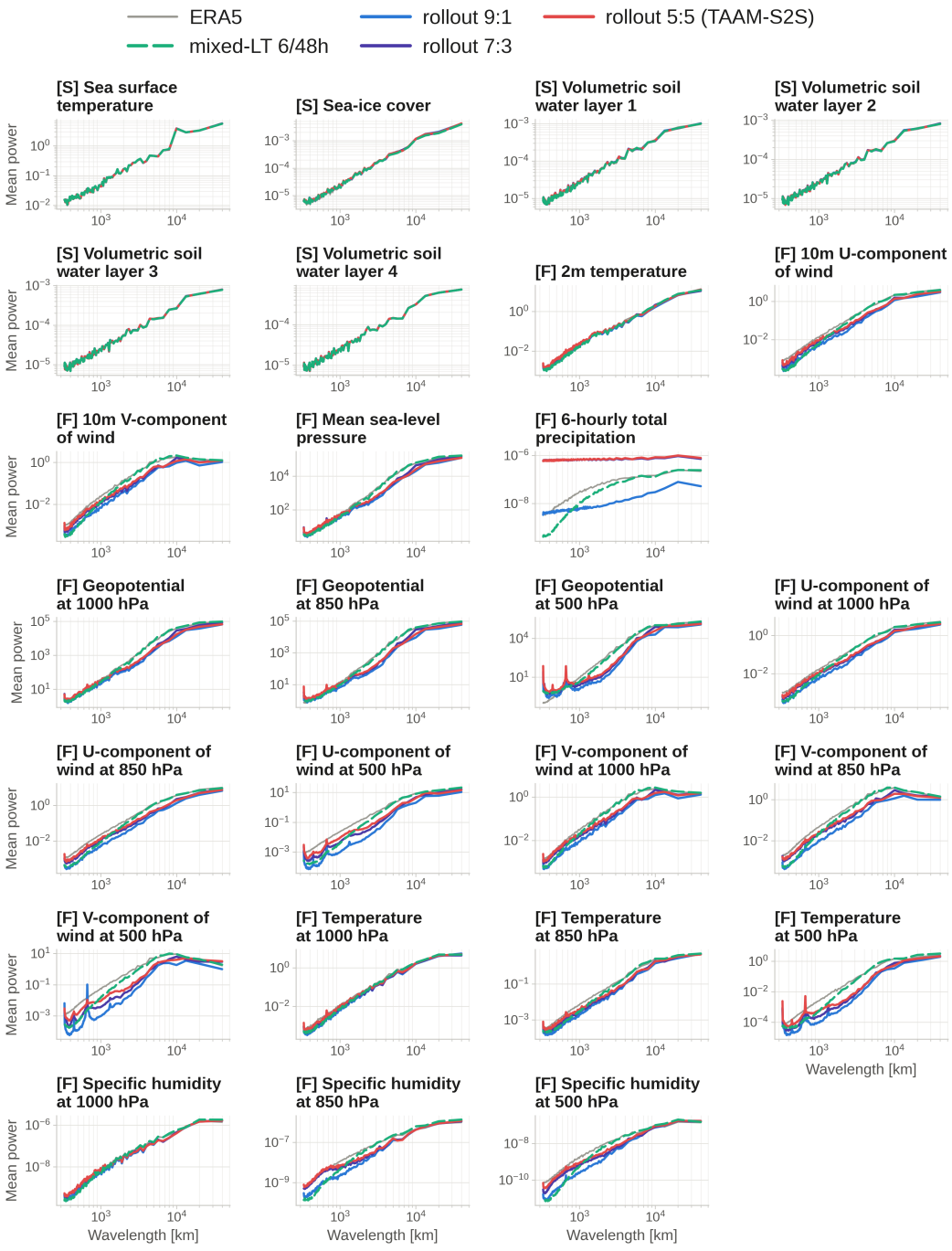

Figure 23: Power spectra of variables predicted at 1008h (42 days) lead time by the rollout-finetuned models and the mixed-LT 6/48h residual model.

K BENCHMARK AGAINST STATE-OF-THE-ART MODELS

Figures 24 and 25 depict the RMSE and LSE of the TAAM-S2S model compared to AIFS v1.1, Aurora v1.5 and DLESyM. TAAM-S2S consistently matches or outperforms the other models on slow variables, while slightly lagging behind in accuracy on fast variables at mid-range time horizons.

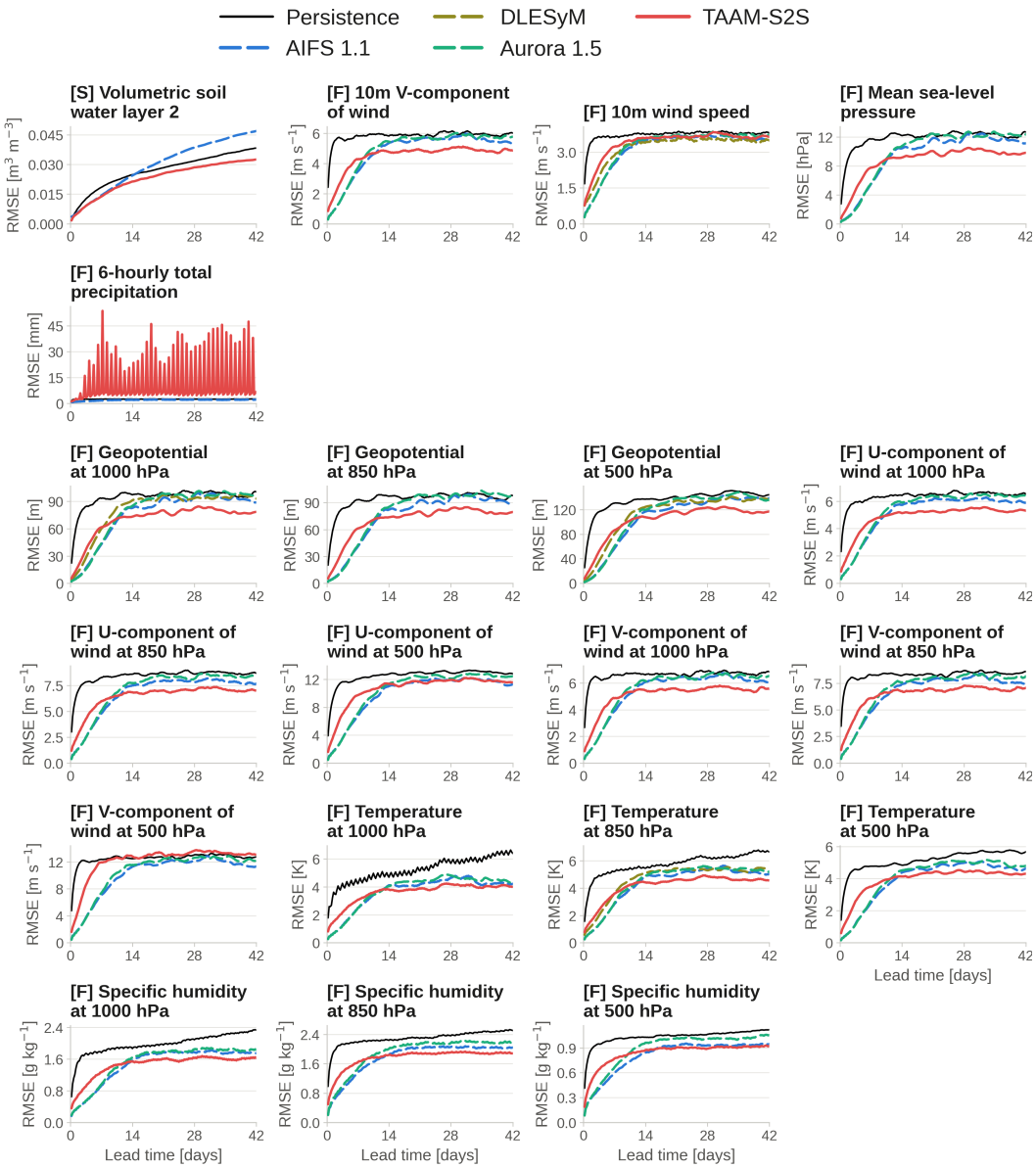

Figure 24: RMSE (↓) as a function of lead time for all the remaining variables omitted from Figure 6. [F] denotes a fast variable and [S] a slow variable.

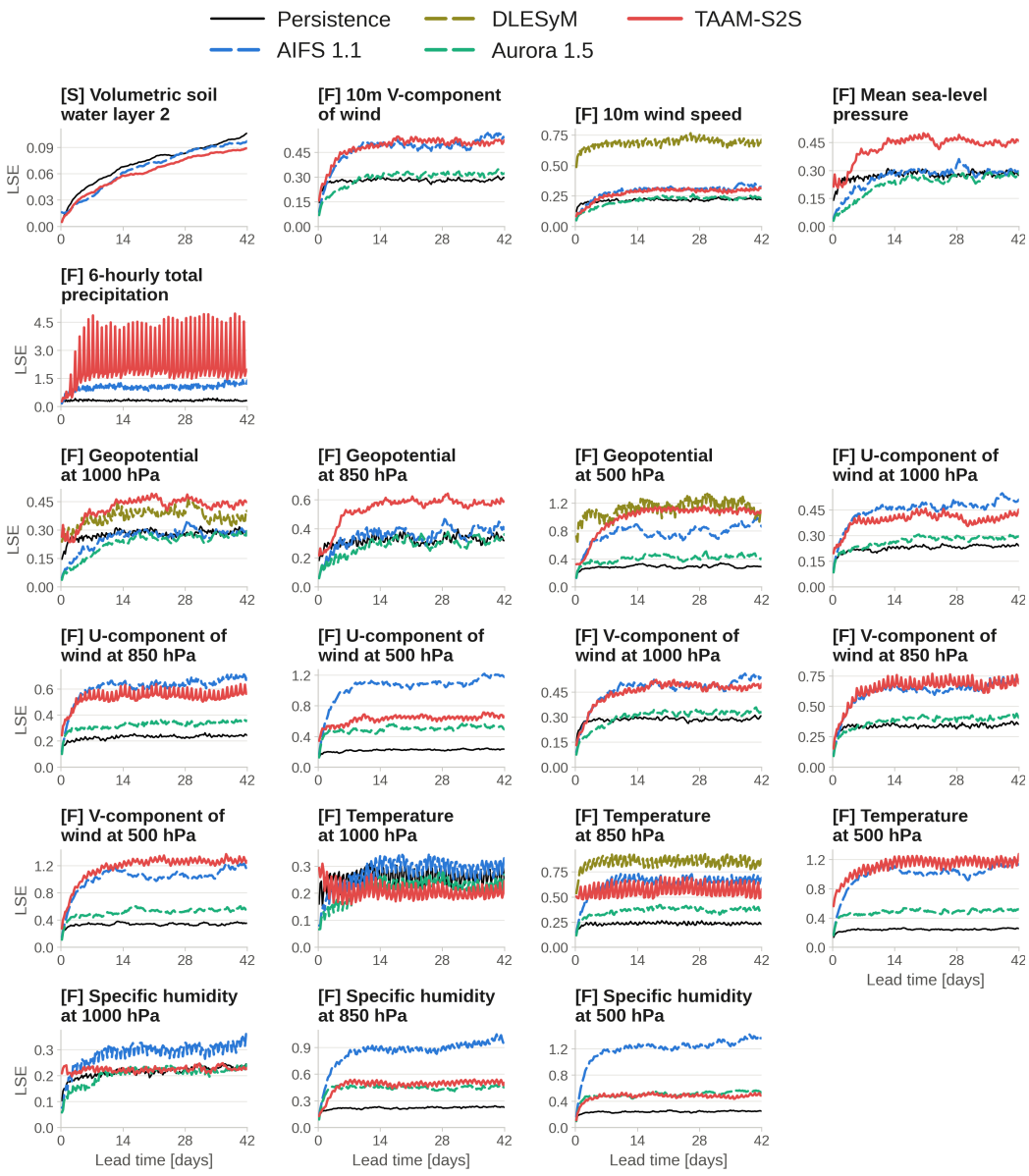

Figure 25: LSE (↓) as a function of lead time for all the remaining variables omitted from Figure 6. [F] denotes a fast variable and [S] a slow variable.