MARCHUK-S2S: COMPACT LATENT FLOW MATCHING FOR SIX-HOURLY SUBSEASONAL ENSEMBLE WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Subseasonal forecasts two to six weeks ahead must draw predictability from slowly evolving modes such as the Madden–Julian Oscillation (MJO) after the memory of the initial state has faded, and are therefore issued as large ensembles. Diffusion- and flow-based weather models generate sharp six-hourly ensembles but have been evaluated mainly in the medium range. We introduce Marchuk-S2S, a 276M-parameter Transformer trained with conditional flow matching in the latent space of a frozen pretrained weather autoencoder. Conditioned on one atmospheric state and time-of-year labels, it jointly samples multi-day chunks of six-hourly states; a single checkpoint trained on 1- to 8-day chunks is composed autoregressively to 48 days, without lead-time-specific networks or gridded climatology. A 50-member, 48-day ensemble takes 9.7 minutes on one H100 GPU, $7.6\times$ faster than LaDCast. Over 2019–2021, Marchuk-S2S is competitive with FuXi-S2S and the 51-member ECMWF ensemble in weekly anomaly correlation and MJO skill. Training on six-hourly states and aggregating to daily means lowers CRPS on all five evaluated fields compared with training on daily means. Treating each call as a stochastic transition, we show that composing two chunks keeps the ensemble scale of one direct chunk of twice the length and lowers its CRPS, without a consistency loss.

1 INTRODUCTION

Subseasonal-to-seasonal (S2S) forecasts, covering lead times of roughly two to six weeks, support early warning and resource planning in sectors such as emergency response, energy, water management, and agriculture (White et al., 2022). Their importance is growing as climate change increases the likelihood and severity of many types of extreme weather (Seneviratne et al., 2021). At the same time, S2S prediction is notoriously hard: initial-state errors amplify under chaotic atmospheric dynamics (Lorenz, 1963; Zhang et al., 2019), skill degrades rapidly beyond two weeks, and many targets approach climatological benchmarks (Kieu, 2024). At these lead times a single deterministic trajectory cannot represent the growing uncertainty, so forecasts must be probabilistic and are typically issued as large ensembles (Slingo & Palmer, 2011).

Deep learning has transformed medium-range weather prediction. Deterministic models such as Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), and Aurora (Bodnar et al., 2025) match or outperform the operational IFS-HRES on many headline scores, and the diffusion-based GenCast (Price et al., 2025) outperforms the IFS-ENS ensemble on most verification targets. LaDCast (Zhuang & Duraisamy, 2025) generates such ensembles far more cheaply in the compressed latent space of a pretrained deep autoencoder, but only to 15 days. Learned systems also reach subseasonal leads. Weyn et al. (2021) produced 320-member, six-week ensembles at six-hourly resolution with a set of CNNs on a cubed sphere; FengWu-W2S (Ling et al., 2024) rolls out six-hourly forecasts to 42 days with ocean–atmosphere–land coupling and ensemble perturbations; and AIFS-CRPS (Lang et al., 2026), a stochastic autoregressive model trained with a CRPS loss on rollouts of at most 72 hours, matches or exceeds the IFS ensemble through 46 days, a line extended by AIFS-SUBS (Schloer et al., 2026) with a 24-hour step. FuXi-S2S (Chen et al., 2024) and TianQuan-S2S (Li et al., 2026) target daily means, and OmniCast (Nguyen et al., 2025) jointly generates 44 daily states with masked latent diffusion. In these systems each network call either advances the state

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.

by one fixed step or produces daily states, while diffusion- and flow-based forecasters of six-hourly states, such as GenCast, LaDCast, Swift (Stock et al., 2026), and Tyche (Xu et al., 2026), have been scored mainly in the medium range. Long rollouts of medium-range models often become less skillful than climatology after two weeks (Nathaniel et al., 2024).

This leads to the question studied in this paper: *can a compact latent generative model that jointly samples multi-day chunks of six-hourly states, with one network for all lead times and no gridded climatology, be composed to subseasonal leads while remaining competitive with dedicated S2S systems?* This requires a model that stays stable and skillful when its own samples are fed back far beyond the training horizons, and that is cheap enough to generate many long ensembles: a 50-member, 48-day six-hourly ensemble has 192 output states per member.

We introduce **Marchuk-S2S**, a 276M-parameter Transformer trained with conditional flow matching (Lipman et al., 2023) in the latent space of the frozen LaDCast autoencoder. Each model call conditions on a single atmospheric state and on the time-of-year labels of the source and target times, and jointly samples a multi-day chunk of six-hourly latent states. One checkpoint is trained on chunks of 1, 2, 4, and 8 days and is composed autoregressively in four-day chunks to 48 days. Our contributions are:

- **A latent flow-matching forecaster for subseasonal ensembles.** Marchuk-S2S combines a frozen, highly compressed pretrained latent space, conditional flow matching, and variable-horizon joint generation of multi-day six-hourly chunks. One set of weights covers the medium range to 48 days, and a 50-member, 192-step ensemble takes 9.7 minutes on one H100 GPU, $7.6\times$ less than the 1.6B-parameter LaDCast.
- **Six-hourly training improves daily forecasts.** Aggregating six-hourly forecasts to daily means gives lower CRPS on all five evaluated fields than training the same model directly on daily means. Weeks 3–4 anomaly correlation is higher for U_{850} (0.18 to 0.26) and precipitation, similar for T_{850}, and 0.01–0.02 lower for Z_{500} and T2M.
- **Separated conditioning and variable-horizon training.** Distinct mechanisms encode grid position, sequence position, time of year, and flow time. Ablations support the positional-encoding choice, variable-horizon training, and latent-space CRPS fine-tuning slightly lowers day-15 CRPS.
- **A stochastic-transition view of long rollouts.** We measure how far direct and composed transitions are from the Chapman–Kolmogorov relation, relative to the sampling floor of two independent ensembles. Without a consistency loss, composing two chunks keeps a similar ensemble scale and lowers CRPS relative to one direct chunk of twice the length.
- **A medium-range and subseasonal evaluation.** Over the held-out period 2019–2021 and at nearly equal ensemble size (50 versus 51 members), Marchuk-S2S is competitive with FuXi-S2S and the ECMWF S2S ensemble in weekly anomaly correlation and Madden–Julian Oscillation (MJO) skill, and its spatial spectra stay close to ERA5 over 48-day rollouts.

2 Related Work

Machine-learning weather forecasting. Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FuXi (Chen et al., 2023), and Aurora (Bodnar et al., 2025) are strong deterministic medium-range forecasters trained on ERA5 (Hersbach et al., 2020) and commonly compared through WeatherBench-2 (Rasp et al., 2024). Probabilistic models instead learn a conditional distribution over future states. GenCast (Price et al., 2025) applies autoregressive diffusion on the physical grid, whereas LaDCast (Zhuang & Duraisamy, 2025) performs diffusion in the latent space of a pretrained deep-compression autoencoder (DC-AE) and produces six-hourly ensembles to 15 days. ArchesWeatherGen (Couairon et al., 2026) uses flow matching to map a deterministic forecast toward the distribution of ERA5 states, and AIFS-CRPS (Lang et al., 2026) trains a stochastic autoregressive model directly with an almost-fair CRPS objective. The consistency model Swift (Stock et al., 2026) and the one-step conditional flow Tyche (Xu et al., 2026) reduce sampling cost; both are scored in the medium range, and Swift also remains stable in 75-day rollouts. Their notion of consistency concerns the generative path; it differs from the consistency between direct and composed transitions in *physical* forecast time studied here.

Subseasonal forecasting. Operational S2S prediction relies on large numerical ensembles such as those collected in the S2S database (Vitart et al., 2017). Learned six-hourly subseasonal ensembles already exist. Weyn et al. (2021) applied 32 CNNs with different weights to 10 perturbed initial states, giving 320-member six-week forecasts that approach ECMWF over land at weeks 4–6. FengWu-W2S (Ling et al., 2024) rolls out a coupled ocean–atmosphere–land model with initial-condition and model perturbations to 42 days. AIFS-CRPS is evaluated through 46 days and matches or exceeds the IFS ensemble, and AIFS-SUBS (Schloer et al., 2026) adapts it with a 24-hour step and a stricter verification protocol. FuXi-S2S (Chen et al., 2024) and TianQuan-S2S (Li et al., 2026) forecast daily means to 42 and 45 days, the latter with climatological conditioning and separate models for successive lead-time segments, and OmniCast (Nguyen et al., 2025) jointly generates 44 daily states with masked latent diffusion.

Position of this work. Six-hourly subseasonal ensembles and latent generative S2S forecasting thus both exist. Our contribution is their combination: conditional flow matching in a frozen, $\approx 64\times$ compressed pretrained latent space, with one network that jointly samples multi-day six-hourly chunks, is trained on several chunk horizons, and is composed to 48 days without gridded climatology. We compare directly with ECMWF and FuXi-S2S; comparisons with the other systems, which differ in grid, period, and verification, remain open (Appendix A.1).

3 METHOD

3.1 PROBLEM SETUP AND DATA

We use ERA5 (Hersbach et al., 2020) on the WeatherBench-2 1.5° grid (240×121) (Rasp et al., 2024), with states separated by $\Delta = 6\,\mathrm{h}$. Following LaDCast, each state x_t contains six atmospheric variables (geopotential, temperature, zonal and meridional wind, specific humidity, and vertical velocity) at 13 pressure levels and six surface variables (2-m temperature, 10-m zonal and meridional wind, mean sea-level pressure, sea-surface temperature, and 6-hour accumulated precipitation), for 84 dynamic channels. Five static geographical fields enter only the frozen autoencoder (Appendix A.2). The model and the frozen DC-AE (Zhuang & Duraisamy, 2025) are trained on 1979–2017; 2018 is reserved for ablations and 2019–2021 for testing.

For a source time t, the *context* is always the single state x_t, and the *target* is the trajectory $\mathbf{x}_{t,1:L} = (x_{t+\Delta}, \ldots, x_{t+L\Delta})$. We call $h = L\Delta$ the *chunk horizon*. The *rollout horizon* H is the total lead time obtained by composing chunks; we use $H = 15$ days for medium-range evaluation and $H = 48$ days for subseasonal forecasts.

3.2 LATENT CONDITIONAL FLOW MATCHING

Let $\mathcal{E}$ be the frozen LaDCast DC-AE encoder, which provides about $64\times$ spatial compression, $z_t = \mathcal{E}(x_t)$ the latent context, and $\mathbf{z}^\star = \mathcal{E}(\mathbf{x}_{t,1:L})$ the clean latent target trajectory. The conditioning is $c = (z_t, \tau_{t:t+h})$, where $\tau_{t:t+h}$ contains time-of-year embeddings of the source time and of all target valid times. We use conditional flow matching on a straight path (Lipman et al., 2023): for flow time $s \sim \mathcal{U}[0,1]$ and $\varepsilon \sim \mathcal{N}(0, I)$ with the shape of the target, the network v_θ regresses the velocity of $\mathbf{z}_s = (1-s)\varepsilon + s\mathbf{z}^\star$,

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{s,\varepsilon,\mathbf{z}^\star} \left\| v_\theta(\mathbf{z}_s, s; c) - (\mathbf{z}^\star - \varepsilon) \right\|_2^2. \tag{1}$$

All L future states are generated jointly, so temporal coherence within a chunk is learned rather than imposed by recursion. At inference, integrating the learned velocity field from a fresh Gaussian sample gives one latent trajectory, which the frozen decoder maps to weather fields (Figure 4 in Appendix A.5). The complete sampler defines a random map

$$\widehat{\mathbf{z}}_t^{(h)} = G_\theta^{(h)}\big(z_t, \tau_{t:t+h}, \varepsilon_t\big), \qquad \varepsilon_t \sim \mathcal{N}(0, I). \tag{2}$$

Integration is deterministic given ε_t, and resampling ε_t produces the ensemble. The flow time s is an auxiliary generative coordinate distinct from physical forecast time: stochasticity in physical time arises from sampling and composing conditional trajectory transitions, not from an atmospheric-time stochastic differential equation.

3.3 Architecture and Temporal Conditioning

The backbone has 276M parameters, excluding the frozen DC-AE, and consists of 32 DiT blocks (Peebles & Xie, 2023) with the parameter-efficient AdaLN–LoRA modulation of W.A.L.T. (Gupta et al., 2024): a single modulation MLP is shared across blocks and specialized by lightweight block-specific low-rank adapters (Appendix A.5). Marchuk-S2S represents four distinct notions of position and time with separate mechanisms. *Spatial position* on the fixed latent grid is encoded by learned 2-D embeddings instead of LaDCast's spherical GeoRoPE prior. *Sequence position* within a generated chunk is encoded by 1-D temporal rotary embeddings (Su et al., 2024), which preserve ordering across variable-length targets. *Time of year* enters through cross-attention: each source or target timestamp is represented by its *month:day:hour* label and mapped to one of $366 \times 4 = 1{,}464$ learned embeddings. The label supplies no gridded climatology; any seasonal information must be learned through the embeddings, and the forecast must still be inferred from the initial state. Its role is to let a single forecasting window serve every initialization date: each call knows where in the annual cycle its source and target valid times fall, so the model can account for the seasonal shift that occurs within a multi-week forecast. Finally, *flow time* s controls the AdaLN–LoRA modulation.

3.4 Variable-Horizon Training and Autoregressive Composition

Variable-horizon training (VHT) samples $h \in \{1, 2, 4, 8\}$ days uniformly for each example, corresponding to $L \in \{4, 8, 16, 32\}$ target states, while the six-hourly spacing and the single-state context stay fixed. One checkpoint therefore learns four trajectory distributions through one conditioning interface; parameter sharing couples the horizons, but no loss explicitly matches them.

At inference, the observed initial state is encoded once. Each ensemble member is then rolled out independently: the final latent state of its previous chunk becomes the next context, the time-of-year embeddings are updated, and a fresh Gaussian sample is drawn. Intermediate states are forecast outputs, not additional context. We use four-day chunks of 16 states, selected by the ablation in Section 4.5, and integrate each chunk with 10 explicit Euler steps. A 48-day forecast thus takes 12 calls and yields 192 six-hourly states per member.

Passing only the current latent state between calls makes the rollout first-order in the latent representation. This is a modeling closure, not a claim that the resolved atmosphere is exactly Markov. Let $p^{(h)}_{\theta,t}(z' \mid z)$ denote the endpoint law of an h-day transition starting at time t. For states z_0, z_1, z_2, exact compatibility of direct and composed transitions would require the Chapman–Kolmogorov relation

$$p^{(h_1+h_2)}_{\theta,t}(z_2 \mid z_0) = \int p^{(h_2)}_{\theta,t+h_1}(z_2 \mid z_1)\, p^{(h_1)}_{\theta,t}(z_1 \mid z_0)\, \mathrm{d}z_1, \tag{3}$$

where the time indices reflect time-of-year conditioning. Long forecasts always use the composed path and do not require exact equality; equation 3 is a reference criterion that we test empirically in Section 4.4.

3.5 Latent-Space CRPS Fine-Tuning

Flow matching fits the marginal trajectory distribution but does not directly target ensemble calibration. After pretraining, we therefore fine-tune on $\mathcal{L}_{\mathrm{FM}} + \lambda\, \mathcal{L}_{\mathrm{CRPS}}$ with $\lambda = 0.1$: the flow-matching loss equation 1 remains the main objective, and $\mathcal{L}_{\mathrm{CRPS}}$ is the fair finite-ensemble CRPS (Ferro, 2014),

$$\mathcal{L}_{\mathrm{CRPS}}(\theta) = \frac{1}{M} \sum_{m=1}^{M} \left\| \widehat{\mathbf{z}}_m - \mathbf{z}^\star \right\|_1 - \frac{1}{M(M-1)} \sum_{m<n} \left\| \widehat{\mathbf{z}}_m - \widehat{\mathbf{z}}_n \right\|_1, \tag{4}$$

where $\widehat{\mathbf{z}}_m$ are $M = 50$ members sampled from independent noise and gradients are backpropagated through the sampler. The first term penalizes error and the second rewards spread without the finite-M bias of the standard estimator. Fine-tuning runs for 2,000 steps with learning rate 10^{-5} in latent space; all reported scores are computed after decoding.

Inference cost. Per output state, Marchuk-S2S is $7.6\times$ cheaper than LaDCast-1.6B and $2.4\times$ cheaper than FuXi-S2S, which is $1.7\times$ cheaper per forecast day because it emits one daily state (Appendix Table 7).

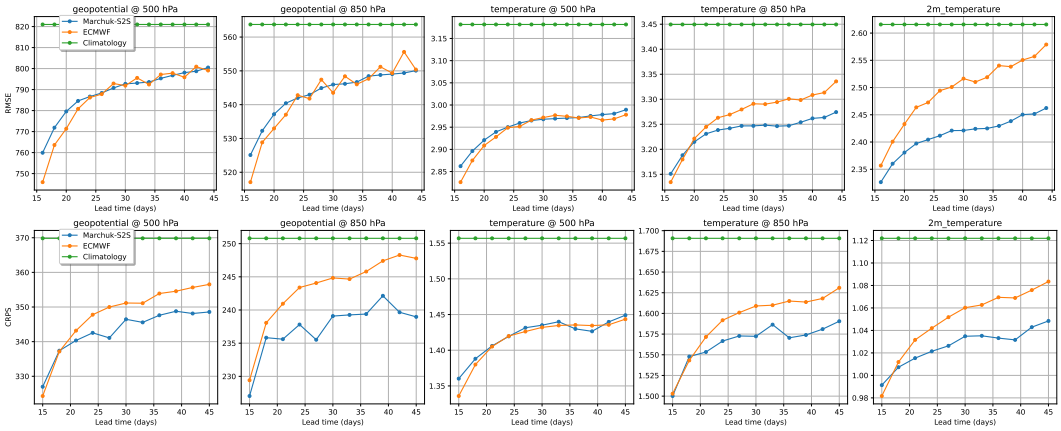

Figure 1: **Subseasonal skill on matched 2019–2021 initializations.** Ensemble-mean RMSE (top) and CRPS (bottom) of daily means from days 15 to 45 for Marchuk-S2S (50 members), the ECMWF S2S ensemble (51 members), and climatology, for Z_{500}, Z_{850}, T_{500}, T_{850}, and T2M. Every Marchuk-S2S forecast is obtained by repeated composition of four-day transitions.

4 Experiments

We evaluate decoded Marchuk-S2S forecasts in five steps: medium-range and subseasonal skill against operational and learned baselines (Section 4.1), subseasonal anomaly and MJO skill (Section 4.2), the effect of six-hourly training (Section 4.3), long-rollout structure and direct–composed transition compatibility (Section 4.4), and controlled ablations (Section 4.5). All main results and the CRPS fine-tuning ablation use $M = 50$ members; all other ablations use $M = 20$.

4.1 Evaluation Protocol and Forecast Skill

Protocol. Following WeatherBench-2 (Rasp et al., 2024), we report latitude-weighted ensemble-mean RMSE and ensemble CRPS on the 1.5° ERA5 grid. Medium-range scores follow the WeatherBench-2 test year and use 732 initializations in 2020 (00 and 12 UTC on each day). All subseasonal evaluations cover the held-out period 2019–2021. For the ECMWF S2S ensemble (Vitart et al., 2017) we use its twice-weekly initialization dates in this period, with the same grid and ERA5 targets for all systems. The ECMWF S2S ensemble and FuXi-S2S have 51 members and Marchuk-S2S has 50, so probabilistic scores are compared at nearly equal ensemble size. For comparisons with daily systems, the four six-hourly states of each day (00–18 UTC) are averaged and the four six-hour precipitation accumulations are summed, for both Marchuk-S2S and ERA5. Ablations use 2018 initializations, outside all test periods, so design choices were not selected on test forecasts.

Medium range. Over 15 days (Appendix A.9), Marchuk-S2S is compared with LaDCast-1.6B, GraphCast, Pangu-Weather, IFS-ENS, IFS-HRES, and climatology. Its RMSE and CRPS are somewhat higher than those of LaDCast-1.6B during the first week on several upper-air fields and converge to them by days 10–15, where Marchuk-S2S is slightly better on some fields such as Z_{500} and T2M CRPS. IFS-ENS remains the strongest baseline, with a variable-dependent gap. Skill close to LaDCast-1.6B thus comes with a $5.8\times$ smaller backbone.

Subseasonal range. Figure 1 compares Marchuk-S2S with the ECMWF S2S ensemble from days 15 to 45. ECMWF is more accurate at the start of this window for geopotential and T_{500}, but Marchuk-S2S has lower CRPS over much of the range for Z_{500}, Z_{850}, T_{850}, and T2M and remains close on T_{500}. At late leads its RMSE is lower for T_{850} and T2M and comparable on the other fields. All errors stay below those of climatology through day 45. ECMWF is verified without bias correction and starts from its own analysis, so part of its error growth, e.g. for T2M, may be model drift, and raw scores reward reproducing the seasonal cycle; anomaly-based scores (Section 4.2) are therefore our primary measure.

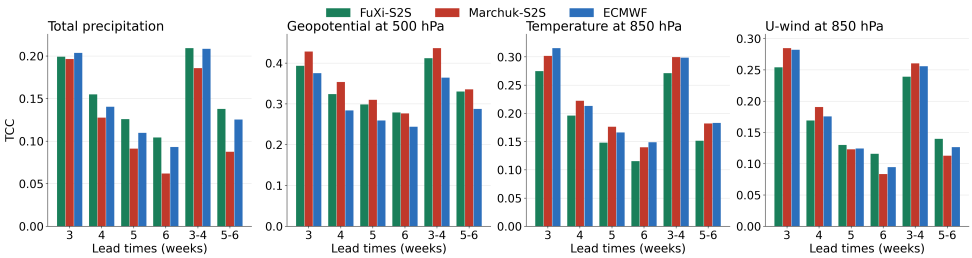

Figure 2: **Weekly anomaly correlation on 2019–2021 forecasts.** Globally averaged, latitude-weighted temporal anomaly correlation (TCC) of the ensemble mean for weeks 3–6 and the aggregated weeks 3–4 and 5–6, for the ECMWF S2S ensemble, FuXi-S2S, and Marchuk-S2S on the matched initialization dates of Section 4.1. Variables are 24-h total precipitation, Z_{500}, T_{850}, and U_{850}; anomalies are relative to the 1990–2017 ERA5 day-of-year climatology.

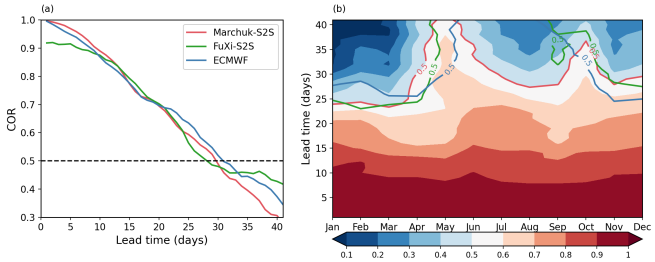

Figure 3: **MJO forecast skill.** Bivariate correlation (COR) of a precipitation-based RMM-like index for the ensemble means of ECMWF (blue), FuXi-S2S (green), and Marchuk-S2S (red), averaged over the matched 2019–2021 initializations. (a) COR versus lead time; the dashed line marks the conventional skill threshold $\mathrm{COR} = 0.5$. (b) COR by initialization month and lead time, with the 0.5 contour of each system.

4.2 Subseasonal Anomaly and MJO Skill

Temporal anomaly correlation. Forecasts and ERA5 are expressed as anomalies relative to a day-of-year ERA5 climatology computed over 1990–2017, before the test period, and the temporal Pearson correlation over initializations is averaged globally with latitude weights. Because the climatology is removed, TCC measures skill beyond the mean seasonal cycle. Figure 2 compares Marchuk-S2S with FuXi-S2S and the ECMWF S2S ensemble over 2019–2021. Marchuk-S2S has the highest Z_{500} TCC in most weekly windows and is competitive on T_{850}. For U_{850} it is highest in week 4 and within about 0.005 of ECMWF in week 3 and weeks 3–4, differences too small to separate without confidence intervals. Its precipitation TCC is the lowest of the three systems from week 3 and decreases faster than that of both baselines, making precipitation the clearest subseasonal weakness of the model.

Madden–Julian Oscillation. The MJO, an eastward-propagating envelope of tropical convection on 30–60-day timescales (Madden & Julian, 1971; 1972), is a key source of subseasonal predictability. We compute an RMM-like index (Wheeler & Hendon, 2004) from U_{850}, U_{200}, and total precipitation, which replaces outgoing longwave radiation because the latter is not a model variable, and apply the same index to all three systems and ERA5 (Appendix A.8). In Figure 3, Marchuk-S2S and ECMWF are close during the first ten days, the three systems are close between days 15 and 30, and all fall below the 0.5 threshold at roughly 30 days. Beyond about 32 days, Marchuk-S2S decays faster than both baselines.

4.3 Effect of Six-Hourly Training on Daily Skill

Daily-aggregate scores alone do not show what the six-hourly formulation contributes, so we train a daily counterpart of Marchuk-S2S that is identical except that its targets are daily means, with

Table 1: **Six-hourly versus daily-mean training**, scored on daily means over 2019–2021 ($M = 50$). Both models share the architecture, conditioning, variable-horizon schedule, and number of training iterations; they differ in the temporal resolution of the training targets, so the six-hourly model receives four times more target frames per example. RMSE and CRPS are averaged over days 15–42; TCC is for weeks 3–4. Bold: better of the two.

Metric	Training targets	Z_{500}	T_{850}	U_{850}	T2M	$\mathrm{TP}_{24\mathrm{h}}$
RMSE ↓	Daily means	**786.7**	3.249	5.209	2.512	0.00593
	Six-hourly, aggregated (ours)	787.2	**3.234**	**5.200**	**2.410**	**0.00584**
CRPS ↓	Daily means	351.4	1.594	2.701	1.103	0.001868
	Six-hourly, aggregated (ours)	**343.1**	**1.568**	**2.649**	**1.027**	**0.001813**
TCC ↑	Daily means	**0.45**	0.29	0.18	**0.29**	0.15
	Six-hourly, aggregated (ours)	0.44	**0.30**	**0.26**	0.27	**0.19**

24-hour precipitation totals, instead of six-hourly states. Both models are evaluated on the same initializations and daily ERA5 targets, and the six-hourly forecasts are aggregated to daily resolution as in Section 4.1. Aggregating after training improves probabilistic skill most consistently (Table 1): CRPS falls on all five fields, by 1.6–6.9%. Weeks 3–4 TCC is higher for U_{850} (0.18 to 0.26) and precipitation (0.15 to 0.19), similar for T_{850} (0.29 and 0.30), and 0.01–0.02 lower for Z_{500} and T2M. Ensemble-mean RMSE falls most for T2M (4.1%) and precipitation (1.5%), two fields with a strong diurnal cycle, changes by less than 0.5% for T_{850} and U_{850}, and is unchanged for Z_{500}. This pattern is consistent with the hypothesis that six-hourly targets also supervise the diurnal cycle and short-timescale dynamics, which daily averaging removes before training; for T2M, the gain appears in absolute errors rather than in anomaly correlation. The six-hourly model also receives more supervision per example, and daily means, especially 24-h precipitation totals, lie partly outside the training distribution of the DC-AE, which saw instantaneous states.

4.4 Long-Rollout Structure and Transition Compatibility

Structural diagnostics. Following Swift (Stock et al., 2026), Figure 9 (Appendix A.10) checks for the smoothing that scalar scores can hide. In one 48-day case, tropical U_{850} anomalies keep their main longitude–time organization, though smoother than ERA5 at times; single-member Q_{700} fields remain structured at day 45; and zonal spectra averaged over 50 initializations stay close to ERA5 at all evaluated leads.

Direct versus composed transitions. To probe the first-order state interface of Section 3.4, we use the shared checkpoint to compare direct and composed transitions that start from the same encoded ERA5 state and end at the same valid time: 1+1 versus 2 days, 2+2 versus 4 days, and 4+4 versus 8 days, on 50 initializations from 2018 with $M = 50$ members, averaging over T_{500}, Z_{500}, MSLP, 10m-UW, and T2M. Each composed member continues from its own intermediate endpoint with fresh noise. We report a mean discrepancy $\widetilde{D}_{\mathrm{mean}}$, the distance between the two ensemble means relative to their forecast error; a distributional discrepancy $\widetilde{D}_{\mathrm{dist}}$, an energy distance relative to ensemble width; and the composed-to-direct ratios of spread, R_{spread}, and CRPS, R_{CRPS} (definitions in Appendix A.6). Two independent ensembles from the same calibrated law already give $\widetilde{D}_{\mathrm{mean}} \approx \sqrt{2/(M+1)} = 0.20$ and $\widetilde{D}_{\mathrm{dist}} \approx 2/M = 0.04$; this floor is the reference.

In Table 2, $\widetilde{D}_{\mathrm{mean}}$ (0.38–0.49) is about twice its floor, so sampling noise explains much of it, but $\widetilde{D}_{\mathrm{dist}}$ (0.13–0.18) is three to four times its floor: direct and composed endpoint laws differ. Spread ratios of 0.91–0.94 indicate a similar scale, and composition lowers CRPS by 3–9%, without a consistency loss. This compares two chunks with one of twice the length; over full rollouts, 2–4-day chunks are best (Table 3c).

Table 2: **Direct–composed transition compatibility** on 50 held-out 2018 initializations ($M = 50$; mean $\pm$ standard error; averaged over five fields). Two independent ensembles from the same law give $\widetilde{D}_{\mathrm{mean}} \approx 0.20$, $\widetilde{D}_{\mathrm{dist}} \approx 0.04$, and unit ratios; $R < 1$ means the composed path has smaller spread or CRPS.

Comparison	$\widetilde{D}_{\mathrm{mean}} \downarrow$	$\widetilde{D}_{\mathrm{dist}} \downarrow$	R_{spread}	R_{CRPS}
1+1 vs. 2 days	0.383 ± 0.004	0.128 ± 0.002	0.922 ± 0.001	0.971 ± 0.003
2+2 vs. 4 days	0.488 ± 0.007	0.177 ± 0.004	0.905 ± 0.003	0.910 ± 0.005
4+4 vs. 8 days	0.450 ± 0.008	0.157 ± 0.004	0.936 ± 0.004	0.934 ± 0.007

Table 3: **Day-15 ablations** (ensemble-mean RMSE, $M = 20$, 2018 initializations). Each block varies one factor from its own base configuration; compare values only within a block. All models condition on one state. Bold: best; underlined: second best in block (b).

Configuration	T_{500}	Z_{500}	MSLP	10m-UW	T2M
(a) Positional encoding					
3-D RoPE	3.23	829.83	730.81	3.98	2.66
GeoRoPE (LaDCast)	3.20	821.28	729.82	3.97	2.67
Temporal RoPE + learned 2-D (ours)	**3.18**	**808.11**	**719.79**	**3.94**	**2.65**
(b) Training chunk horizon (separate fixed-horizon models vs. one VHT model)					
6 hours	3.21	820.05	727.83	4.02	2.68
12 hours	3.15	806.22	714.67	3.93	2.63
1 day	3.17	809.98	715.94	3.95	2.66
2 days	<u>3.14</u>	<u>804.05</u>	<u>713.98</u>	3.93	<u>2.61</u>
4 days	3.15	815.62	728.41	**3.89**	**2.57**
8 days	3.20	826.77	741.44	4.11	2.65
VHT, 1/2/4/8 days (ours)	**3.13**	**803.90**	**713.87**	<u>3.92</u>	**2.57**
(c) Inference chunk horizon (fixed VHT checkpoint)					
1 day	3.15	808.56	717.27	**3.92**	2.68
2 days	**3.13**	**800.60**	715.86	**3.92**	2.59
4 days (ours)	**3.13**	803.90	**713.87**	**3.92**	**2.57**
8 days	3.14	804.26	716.23	3.93	2.64

4.5 Ablations

Positional encoding. Temporal RoPE with learned spatial embeddings lowers RMSE on all five fields relative to 3-D RoPE and GeoRoPE (Table 3a): on a fixed latent grid, learned spatial identities beat a spherical prior.

Training and inference horizons. Among separately trained fixed-horizon models (Table 3b), two-day chunks are best or near-best on most fields, four-day chunks on 10m-UW and T2M, 12-hour chunks are competitive, and the eight-day model degrades on all fields. VHT is best or tied on four fields and within 0.03 on 10m-UW, with one checkpoint instead of one per horizon. With the VHT checkpoint fixed (Table 3c), two-day inference chunks give the lowest Z_{500} error and four-day chunks the lowest MSLP and T2M errors with half as many calls, so we use four-day chunks.

CRPS fine-tuning. Fine-tuning with equation 4 lowers day-15 CRPS slightly, by 0.1–0.6%, on all five reported fields, including 6-hour precipitation (Appendix Table 6), while ensemble-mean RMSE is broadly preserved: Z_{500} and T2M improve slightly, T_{500} is unchanged, and MSLP and 10m-UW increase marginally. The final configuration and $M = 50$ both lower RMSE, so these values ($Z_{500} \approx 774$) are not comparable with Table 3 (≈ 800).

5 Conclusion and Limitations

Marchuk-S2S shows that a compact latent generative model that jointly samples multi-day chunks of six-hourly states, with one network for all lead times and no gridded climatology, can be composed to 48 days and remain competitive with dedicated subseasonal-to-seasonal systems on most evaluated fields, with precipitation as the main exception. (i) Combining a frozen, highly compressed pretrained latent space, conditional flow matching, and variable-horizon joint generation, one 276M-parameter checkpoint covers the medium range to 48 days and produces a 50-member, six-hourly, 48-day ensemble in 9.7 minutes on one H100 GPU, $7.6\times$ faster than the 1.6B-parameter LaDCast. (ii) Over the held-out period 2019–2021, Marchuk-S2S is competitive with FuXi-S2S and the ECMWF subseasonal ensemble in weekly anomaly correlation and Madden–Julian Oscillation skill, and its spatial spectra stay close to those of ERA5 over 48-day rollouts. (iii) Aggregating six-hourly forecasts to daily means lowers the daily continuous ranked probability score (CRPS) on all five evaluated fields relative to training the same model on daily means, so learned subseasonal systems that issue daily products may benefit from training at sub-daily resolution. (iv) Separate encodings of spatial position, sequence position, time of year, and flow time, together with variable-horizon training, improve forecast skill at a 15-day lead, and adding a latent-space CRPS term to the flow-matching loss lowers CRPS slightly. (v) Direct and composed transitions produce endpoint distributions that differ by more than the sampling floor of two independent ensembles; nevertheless, composing two chunks keeps an ensemble scale similar to that of one direct chunk of twice the length and lowers its CRPS, without any consistency loss.

Limitations. Precipitation anomaly skill and Madden–Julian Oscillation skill at late leads decay faster than in both baselines, and the subseasonal ensembles are slightly underdispersive (Appendix A.10). At subseasonal leads, FuXi-S2S is the only learned system we compare with directly. The ECMWF raw-field scores are not bias-corrected, and anomalies are computed relative to an observed rather than a model climatology. The evaluation covers three test years without confidence intervals, the ablations are reported only at a 15-day lead, and the value of intraday output to users is not assessed. The Madden–Julian Oscillation index uses precipitation instead of outgoing longwave radiation, which the model does not predict. Analysis and parameter uncertainty are not represented, latent compression may discard small-scale information, and our diagnostics support finite-horizon stability rather than long-run climate fidelity.

Ethics Statement

This work uses only publicly available reanalysis data (ERA5) and published forecast archives; it involves no human subjects or personal data. Improved subseasonal forecasts may support preparedness for hazardous weather, but forecasts can also be over-trusted. Marchuk-S2S has not been evaluated in operational decision-making, and its outputs should not replace official warnings. We report weaknesses such as reduced precipitation skill after week 3 and mild underdispersion so that users can calibrate their trust in the model.

Reproducibility Statement

Section 3 specifies the data, the flow-matching objective, the architecture, variable-horizon training, the composition procedure, and CRPS fine-tuning. Appendix A.2 lists all input and output variables, Appendix A.3 gives the training hyperparameters and compute, Appendix A.5 describes the backbone, and Appendices A.7 and A.8 define the spread–skill ratio and the MJO index. The evaluation follows the public WeatherBench-2 protocol; the matched-date, aggregation, and anomaly procedures are described in Section 4. The frozen autoencoder is the pretrained LaDCast DC-AE (Zhuang & Duraisamy, 2025). The source code, trained model checkpoints, and evaluation scripts will be made publicly available upon acceptance of the paper.

AI Use Statement

In this work, we used generative AI tools for the following tasks: (i) to aid and polish writing, including editing for clarity, consistency, and length; (ii) to draft sections of the paper, which were

written from our own earlier drafts, results, and notes and then revised by the authors; (iii) for retrieval and discovery, such as finding and summarizing related work; and (iv) for research ideation and execution, such as discussing modeling and evaluation choices and assisting with implementation and analysis. We have reviewed all AI-assisted work. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.

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, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Kang Chen, Tao Han, Fenghua Ling, Junchao Gong, Lei Bai, Xinyu Wang, Jing-Jia Luo, Ben Fei, Wenlong Zhang, Xi Chen, et al. The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Communications Earth & Environment*, 6(1):518, 2025b.

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, Zixin Hu, and Yuan Qi. A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 15(1):6425, 2024.

Guillaume Couairon, Renu Singh, Anastase Charantonis, Christian Lessig, and Claire Monteleoni. ArchesWeatherGen: Skillful and compute-efficient probabilistic weather forecasting with machine learning. *Science Advances*, 12(17), 2026.

Nina Effenberger, Nicole Ludwig, and Rachel H White. Mind the (spectral) gap: how the temporal resolution of wind data affects multi-decadal wind power forecasts. *Environmental Research Letters*, 19(1):014015, 2024.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014.

Hayley Fowler, Geert Lenderink, Andreas Prein, Seth Westra, Richard Allan, Nikolina Ban, Renaud Barbero, Peter Berg, Stephen Blenkinsop, Hong Do, Selma Guerreiro, J. Haerter, Elizabeth Kendon, Elizabeth Lewis, Christoph Schaer, Ashish Sharma, Gabriele Villarini, Conrad Wasko, and Xuebin Zhang. Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth & Environment*, 2:107–122, 2021.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, volume 15137 of *Lecture Notes in Computer Science*, pp. 393–411, 2024.

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, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, and J.-N Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Ernest Hovmöller. The trough-and-ridge diagram. *Tellus*, 1(2):62–66, 1949.

Yuan Hu, Lei Chen, Zhibin Wang, and Hao Li. SwinVRNN: A data-driven ensemble forecasting model via learned distribution perturbation. *Journal of Advances in Modeling Earth Systems*, 15 (2), 2023.

Chanh Kieu. Predictability of global ai weather models. *arXiv preprint arXiv:2410.03266*, 2024.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2:18, 2026.

Guowen Li, Xintong Liu, Yang Liu, Mengxuan Chen, Shilei Cao, Xuehe Wang, Juepeng Zheng, Jinxiao Zhang, Haoyuan Liang, Lixian Zhang, Jiuke Wang, Meng Jin, Hong Cheng, and Haohuan Fu. TianQuan-S2S: A subseasonal-to-seasonal global weather model via incorporate climatology state. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hai Lin, Gilbert Brunet, and Jacques Derome. Forecast skill of the Madden–Julian Oscillation in two canadian atmospheric models. *Monthly Weather Review*, 136(11):4130–4149, 2008.

Fenghua Ling, Kang Chen, Jiye Wu, Tao Han, Jing-Jia Luo, Wanli Ouyang, and Lei Bai. Fengwu-w2s: A deep learning model for seamless weather-to-subseasonal forecast of global atmosphere. *arXiv preprint arXiv:2411.10191*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Yang Liu, Zinan Zheng, Jiashun Cheng, Fugee Tsung, Deli Zhao, Yu Rong, and Jia Li. Cirt: Global subseasonal-to-seasonal forecasting with geometry-inspired transformer. In *International Conference on Learning Representations*, volume 2025, pp. 101845–101866, 2025a.

Yang Liu, Zinan Zheng, Yu Rong, Deli Zhao, Hong Cheng, and Jia Li. Equivariant and invariant message passing for global subseasonal-to-seasonal forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1879–1890, 2025b.

Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2): 130–148, 1963.

Roland A. Madden and Paul R. Julian. Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *Journal of the Atmospheric Sciences*, 28(5):702–708, 1971.

Roland A. Madden and Paul R. Julian. Description of global-scale circulation cells in the tropics with a 40–50 day period. *Journal of the Atmospheric Sciences*, 29(6):1109–1123, 1972.

Annarita Mariotti, Paolo M. Ruti, and Michel Rixen. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate and Atmospheric Science*, 1 (1):4, 2018.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. *Advances in Neural Information Processing Systems*, 37:43715–43729, 2024.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Tung Nguyen, Tuan Pham, Troy Arcomano, Rao Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. OmniCast: A masked latent diffusion model for weather forecasting across time scales. *Advances in Neural Information Processing Systems*, 38:172182–172212, 2025.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Kathy Pegion, Ben P. Kirtman, Emily Becker, Dan C. Collins, Emerson LaJoie, Robert Burgman, Ray Bell, Timothy DelSole, Dughong Min, Yuejian Zhu, et al. The subseasonal experiment (SubX): A multimodel subseasonal prediction experiment. *Bulletin of the American Meteorological Society*, 100(10):2043–2060, 2019.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallègue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6), 2024.

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.

Sonia I. Seneviratne, Xuebin Zhang, et al. Weather and climate extreme events in a changing climate. In *Climate Change 2021: The Physical Science Basis*, pp. 1513–1766. Cambridge University Press, 2021.

Julia Slingo and Tim Palmer. Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956): 4751–4767, 2011.

Jason Stock, Troy Arcomano, and Rao Kotamarthi. Swift: An autoregressive consistency model for efficient weather forecasting. *Machine Learning: Earth*, 2(2):025004, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Frédéric Vitart, Claire Ardilouze, Anca Bonet, Anca Brookshaw, Mingyue Chen, Constantin Codorean, Michel Déqué, Laura Ferranti, Elisabetta Fucile, Manuel Fuentes, Harry Hendon, James Hodgson, Hyun-Suk Kang, Arun Kumar, Hai Lin, Guoqiang Liu, Xiongbiao Liu, Piero Malguzzi, Ilias Mallas, Michail Manoussakis, Daniela Mastrangelo, Craig MacLachlan, Paul McLean,

Ayako Minami, Roman Mladek, Tetsuo Nakazawa, Saeed Najm, Yiming Nie, Michel Rixen, Andrew W. Robertson, Paolo Ruti, Caio Sun, Yuhei Takaya, Mikhail Tolstykh, Federica Venuti, Duane Waliser, Steven Woolnough, Tongwen Wu, Dae-Je Won, Hui Xiao, Rustam Zaripov, and Li Zhang. The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, 98(1):163–173, 2017.

Frédéric Vitart, Andrew W. Robertson, Aaron Spring, Florian Pinault, Rok Roškar, W. Cao, S. Bech, A. Bienkowski, N. Caltabiano, E. De Coning, et al. Outcomes of the WMO prize challenge to improve subseasonal to seasonal predictions using artificial intelligence. *Bulletin of the American Meteorological Society*, 103(12), 2022.

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, 2021.

Matthew C. Wheeler and Harry H. Hendon. An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, 132(8):1917–1932, 2004.

Christopher J White, Daniela IV Domeisen, Nachiketa Acharya, Elijah A Adefisan, Michael L Anderson, Stella Aura, Ahmed A Balogun, Douglas Bertram, Sonia Bluhm, David J Brayshaw, et al. Advances in the application and utility of subseasonal-to-seasonal predictions. *Bulletin of the American Meteorological Society*, 103(6), 2022.

Fan Xu, Yuan Gao, Kun Wang, Rui Su, Fenghua Ling, Hao Wu, and Wanli Ouyang. Tyche: One step flow for efficient probabilistic weather forecasting. *arXiv preprint arXiv:2605.06916*, 2026.

Fuqing Zhang, Y Qiang Sun, Linus Magnusson, Roberto Buizza, Shian-Jiann Lin, Jan-Huey Chen, and Kerry Emanuel. What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, 76(4):1077–1091, 2019.

Yilin Zhuang and Karthik Duraisamy. Ladcast: A latent diffusion model for medium-range ensemble weather forecasting. *arXiv preprint arXiv:2506.09193*, 2025.

A Appendix

This appendix gives an extended discussion of related work, implementation details, and additional results. All figures use the same trained Marchuk-S2S model and the evaluation procedures of Section 4.

A.1 Extended Related Work

Table 4: **Learned and numerical systems most relevant to this work.** Step: temporal spacing of the output (6-h or 24-h states, or daily means). Horizon: longest evaluated lead (*48-day rollouts, scored to 45 days). Strategy: how long leads are reached (AR: autoregressive; direct: a separate model per lead-time segment; joint: the whole sequence generated at once; chunked AR: one model composed in multi-day chunks). Clim.: gridded climatology as input. Ensemble: source of ensemble spread (IC: initial conditions). DLWP denotes the system of Weyn et al. (2021).

System	Step	Horizon	Strategy	Clim.	Ensemble
ECMWF S2S	daily	46 d	NWP	no	perturbed IC, physics
DLWP	6 h	42 d	AR	no	32 CNNs $\times$ 10 IC
FengWu-W2S	6 h	42 d	AR	no	perturbed IC, model
AIFS-CRPS	6 h	46 d	AR	no	CRPS-trained noise
AIFS-SUBS	24 h	46 d	AR	no	CRPS-trained noise
FuXi-S2S	daily mean	42 d	AR	no	perturbed features
TianQuan-S2S	daily mean	45 d	direct	yes	noise in every layer
OmniCast	24 h	44 d	joint	no	latent diffusion
LaDCast	6 h	15 d	AR	no	latent diffusion
Marchuk-S2S	6 h	45 d*	chunked AR	no	latent flow

Deterministic learned weather models. FourCastNet (Pathak et al., 2022) was an early global data-driven model with skill comparable to numerical prediction at short leads, and Pangu-Weather (Bi et al., 2023) and GraphCast (Lam et al., 2023) then surpassed the operational IFS-HRES on many medium-range scores. FuXi (Chen et al., 2023) and FengWu (Chen et al., 2025b) extended skillful deterministic forecasts beyond ten days, Aurora (Bodnar et al., 2025) and ClimaX (Nguyen et al., 2023) pretrain foundation models on heterogeneous atmospheric data, and NeuralGCM (Kochkov et al., 2024) combines a differentiable dynamical core with learned physics. These models differ in how they reach long leads. Pangu-Weather aggregates separate 1-, 3-, 6-, and 24-hour models, FuXi cascades models specialized to successive lead-time windows, and ClimaX trains for a set of fixed lead times. All are trained with regression losses, so their long-lead forecasts approach a conditional mean and become increasingly smooth, which is one motivation for generative forecasters at subseasonal leads.

Probabilistic learned weather models. Learned ensembles are obtained either by perturbing a deterministic network or by learning a conditional distribution. Weyn et al. (2021) combine 32 CNNs with different learned weights with 10 perturbed initial states, FengWu-W2S (Ling et al., 2024) perturbs both initial conditions and model parameters, SwinVRNN (Hu et al., 2023) perturbs a learned latent distribution, FuXi-S2S (Chen et al., 2024) perturbs hidden features in a flow-dependent way, and TianQuan-S2S (Li et al., 2026) injects learnable Gaussian noise into every Transformer layer. Generative approaches include GenCast (Price et al., 2025), which applies autoregressive diffusion on the physical grid; ArchesWeatherGen (Couairon et al., 2026), which uses flow matching to correct a deterministic forecast toward the distribution of ERA5 states; and LaDCast (Zhuang & Duraisamy, 2025), which performs diffusion in the latent space of a deep-compression autoencoder (Chen et al., 2025a). A second line trains stochastic models directly with proper scoring rules, such as AIFS-CRPS (Lang et al., 2026) and the functional generative network of Alet et al. (2025). Swift (Stock et al., 2026) and Tyche (Xu et al., 2026) reduce sampling cost with consistency or one-step flow models combined with rollout-based CRPS supervision; both are scored in the medium range, and Swift is shown to remain stable, without a quantitative subseasonal evaluation, in 75-day rollouts. Marchuk-S2S combines a latent generative objective with a fine-tuning stage that adds a CRPS

term to the flow-matching loss (Section 3.5), and uses the frozen LaDCast autoencoder so that long ensembles remain affordable.

Subseasonal prediction. Operational S2S forecasts come from large numerical ensembles, collected for research in the S2S database (Vitart et al., 2017) and the SubX multimodel experiment (Pegion et al., 2019); see Mariotti et al. (2018) and White et al. (2022) for reviews of progress and applications. Skill at these leads is limited by the growth of initial-state errors (Lorenz, 1963; Zhang et al., 2019), and forecasts must be judged against climatology and in terms of anomalies. Earlier work evaluated classical machine-learning models for subseasonal targets (He et al., 2021), and the WMO prize challenge assessed AI methods for improving S2S predictions (Vitart et al., 2022).

Global learned systems now reach six weeks. Weyn et al. (2021) produced 320-member, six-week DLWP ensembles at six-hourly resolution and 1.4° in about three minutes on one GPU; their CRPS was only modestly worse than that of the ECMWF S2S ensemble over land at weeks 4 and 5–6, and worse at shorter leads. FengWu-W2S rolls out six-hourly forecasts to 42 days (168 steps) at 1.4° with explicit ocean–atmosphere–land coupling, and reports MJO skill to 37 days against 32 days for ECMWF. AIFS-CRPS (Lang et al., 2026) is trained on rollouts of at most 72 hours, yet its 46-day, 8-member reforecasts over 2018–2022 match or exceed the IFS ensemble, with the largest gains at weeks 2–3 and some degradation in the tropics at week 6. AIFS-SUBS (Schloer et al., 2026) adapts it to subseasonal leads with a 24-hour step, stratospheric levels, and top-of-atmosphere radiation, holds out 2007–2011 for verification, matches the IFS in probabilistic skill over weeks 2–6, and extends skillful forecasts of the convective MJO component by eight days. FuXi-S2S produces 42-day daily-mean ensembles and reports MJO skill competitive with ECMWF; TianQuan-S2S forecasts daily means to 45 days; CirT (Liu et al., 2025a) and equivariant message passing (Liu et al., 2025b) address global S2S forecasting with geometry-aware architectures; and OmniCast (Nguyen et al., 2025) encodes states with a VAE and jointly generates 44 daily states with masked latent diffusion. ChaosBench (Nathaniel et al., 2024) shows that autoregressive rollouts of medium-range models often lose skill relative to climatology after about two weeks, and Kieu (2024) studies the predictability of global AI weather models. These results motivate the climatology-relative and anomaly-based evaluation of Section 4.2.

Reaching subseasonal lead times. Learned S2S systems follow three strategies. Direct models map the initial state to a distant target without feeding back their own output. TianQuan-S2S trains a separate model for each five-day segment between days 15 and 45 from five days of history, and in its appendix reports that a single autoregressive variant is better below ten days but worse at longer leads. Joint models such as OmniCast generate the whole sequence at once. Autoregressive models apply one step repeatedly and must remain stable under their own errors: six-hourly steps in DLWP, FengWu-W2S, and AIFS-CRPS, and daily steps in FuXi-S2S and AIFS-SUBS, which lengthens the step to limit error accumulation. AIFS-CRPS in particular shows that training on short rollouts can suffice for skill at 46 days. Marchuk-S2S lies between the joint and autoregressive strategies: one network is trained on several chunk horizons (Section 3.4) and composed in four-day chunks, so a 48-day forecast needs only twelve calls. We measure how far direct and composed transitions of this network are from the Chapman–Kolmogorov relation (Section 4.4); this notion of consistency concerns physical forecast time and differs from the consistency of the generative path used by Swift and Tyche.

Climatology and time information. TianQuan-S2S argues that the initial state becomes insufficient at subseasonal leads and fuses a gridded daily climatology, computed over 1979–2016, with the initial state through attention-based patch embedding. It also adds Fourier encodings of time with an annual period and absolute-time embeddings, and its ablations show that climatology improves forecasts beyond about 25 days. Marchuk-S2S takes a lighter route: it receives no gridded climatology and conditions only on the time-of-year labels of the source and target valid times (Section 3.3). The label lets one window serve all initialization dates and account for the seasonal shift within a forecast, while the atmospheric evolution must still be inferred from the initial state. The two approaches are complementary, and combining gridded climatology with dense six-hourly generation is a natural extension.

Temporal resolution. FuXi-S2S and TianQuan-S2S are trained and evaluated on daily means; TianQuan-S2S, for example, averages four six-hourly samples per day. Daily averaging can obscure

the timing and intraday evolution of events that matter for applications, such as wind-power variability (Effenberger et al., 2024) and short-duration rainfall extremes (Fowler et al., 2021). Six-hourly output at subseasonal leads is not new: DLWP, FengWu-W2S, and AIFS-CRPS all roll out six-hourly steps to six weeks, whereas AIFS-SUBS moves to a 24-hour step to limit error accumulation and OmniCast generates 24-hour states. Section 4.3 addresses a related but different question, the temporal resolution of the training targets: for our chunked model, training on six-hourly states and aggregating afterwards gives lower daily CRPS on all evaluated fields, and higher anomaly correlation for U_{850} and precipitation, than training on daily means. Whether this carries over to the choice of step length in single-step autoregressive models remains open.

Evaluation practice. We follow WeatherBench-2 (Rasp et al., 2024) for medium-range RMSE, CRPS, and spread–skill ratio, use the fair ensemble CRPS (Ferro, 2014) for training, and evaluate subseasonal skill with anomaly correlation and an RMM-like MJO index (Wheeler & Hendon, 2004) scored by bivariate correlation (Lin et al., 2008). TianQuan-S2S reports RMSE, ACC, CRPS, spread–mean error, and relative quantile error on 5.625° and 1.40625° grids for 2017–2018. AIFS-SUBS follows a stricter protocol: it scores weekly means, defines anomalies relative to each model's own reforecast climatology with a by-member, other-years construction, reports fair CRPS skill scores with bootstrap confidence intervals, uses an OLR-based RMM index, and complements reforecasts with the real-time AI Weather Quest. Our evaluation differs in three respects: anomalies are relative to an observed climatology, which leaves systematic model biases in the anomalies of every system; no confidence intervals are reported; and precipitation replaces OLR in the MJO index. Differences in grid, test period, ensemble size, and climatology definition also make published numbers hard to compare directly, so our comparisons with ECMWF and FuXi-S2S use matched initializations, the same 1.5° grid, and the same ERA5 targets (Section 4.1). The AIFS-CRPS reforecast period 2018–2022 overlaps our test period, which makes it a natural target for a direct comparison.

What is and is not new. Six-hourly subseasonal ensembles (Weyn et al., 2021; Ling et al., 2024; Lang et al., 2026), stochastic models trained on short rollouts and evaluated to 46 days (Lang et al., 2026), efficient flow-based forecasting (Xu et al., 2026), and latent generative S2S forecasting (Nguyen et al., 2025) all exist. The contribution of Marchuk-S2S is their particular combination: conditional flow matching in a frozen, $\approx 64\times$ compressed pretrained latent space; joint generation of multi-day six-hourly chunks by one network trained on several chunk horizons; composition of these chunks to 48 days without gridded climatology; and the analyses this design enables, namely the direct–composed transition study and the comparison of six-hourly with daily-mean training targets.

A.2 Input and Forecast Variables

Marchuk-S2S predicts six atmospheric variables at 13 pressure levels and six surface variables, giving 84 dynamic channels (Table 5). The frozen LaDCast DC-AE additionally receives five static geographical channels, giving 89 DC-AE input channels; the static fields are inputs only and are not forecast.

A.3 Training Details

The flow-matching model is trained for 120,000 iterations with batch size 8, learning rate 10^{-4}, and an exponential moving average of the parameters with decay 0.9999; training takes about one day on eight NVIDIA H100 GPUs. For each example, the chunk horizon is sampled uniformly from $\{1, 2, 4, 8\}$ days. The resulting model is fine-tuned for 2,000 steps on $\mathcal{L}_{\mathrm{FM}} + 0.1\,\mathcal{L}_{\mathrm{CRPS}}$, with the latent-space CRPS of equation 4 added to the flow-matching loss equation 1, learning rate 10^{-5}, the same batch size, and $M = 50$ independently sampled members per example. Sampling uses 10 explicit Euler steps per chunk. Table 6 reports the effect of fine-tuning at day 15.

A.4 Inference Cost

Table 7 compares the wall-clock cost of 50-member ensembles. Per forecast day, FuXi-S2S needs 7.3 s and Marchuk-S2S 12.1 s; per output state, Marchuk-S2S needs 3.0 s, $2.4\times$ less than FuXi-S2S.

Table 5: Input and forecast variables. Atmospheric variables are given at 13 pressure levels. Static fields are inputs to the frozen DC-AE only.

Type	Variable	Status
Atmosphere	Geopotential	Input / Forecast
Atmosphere	Specific humidity	Input / Forecast
Atmosphere	Air temperature	Input / Forecast
Atmosphere	Zonal wind component	Input / Forecast
Atmosphere	Meridional wind component	Input / Forecast
Atmosphere	Vertical velocity	Input / Forecast
Surface	10-m zonal wind component	Input / Forecast
Surface	10-m meridional wind component	Input / Forecast
Surface	2-m temperature	Input / Forecast
Surface	Mean sea-level pressure	Input / Forecast
Surface	Sea-surface temperature	Input / Forecast
Surface	6-hour accumulated precipitation	Input / Forecast
Static	Land–sea mask	Input
Static	Standard deviation of orography	Input
Static	Angle of sub-gridscale orography	Input
Static	Anisotropy of sub-gridscale orography	Input
Static	Slope of sub-gridscale orography	Input

Table 6: **Latent-space CRPS fine-tuning at day 15** (final configuration, $M = 50$ evaluation members, as in the main results; 2018 initializations, outside all test periods). CRPS decreases slightly (0.1–0.6%) on all fields, while RMSE changes are small and mixed. Values are not comparable with Table 3, which uses reduced base configurations and $M = 20$.

RMSE	T_{500}	Z_{500}	MSLP	10m-UW	T2M
FM baseline	3.00	774.39	**680.50**	**3.725**	2.50
CRPS fine-tuned	3.00	**770.98**	680.96	3.740	**2.49**
CRPS	U_{500}	SH_{500} (10^{-3})	MSLP	T2M	$\mathrm{TP}_{6\mathrm{h}}$ (10^{-3})
FM baseline	4.152	0.380	298.39	1.0810	0.5192
CRPS fine-tuned	**4.149**	**0.378**	**297.61**	**1.0754**	**0.5161**

A.5 Forecasting Backbone

The frozen LaDCast DC-AE was trained by its authors on hourly ERA5 states from 1979–2017 (Zhuang & Duraisamy, 2025), so it has not seen the test years. It maps each state to an $84 \times 15 \times 30$ latent, an 8×8 spatial compression. Figure 4 summarizes training and sampling, and Figure 5 shows the DiT-style backbone that parameterizes the conditional latent velocity field.

A.6 Transition-Compatibility Metrics

For the comparison of Section 4.4, let X_{d} and X_{c} be the direct and composed endpoint ensembles with means $\bar{x}_{\mathrm{d}}, \bar{x}_{\mathrm{c}}$, let y be the ERA5 target, and let d_1 and d_2 denote area-weighted L_1 distance and RMSE. With $W(X) = M^{-2} \sum_{m,n} d_1(x^{(m)}, x^{(n)})$, cross term $C = M^{-2} \sum_{m,n} d_1(x_{\mathrm{d}}^{(m)}, x_{\mathrm{c}}^{(n)})$, energy distance $D_{\mathrm{E}} = 2C - W(X_{\mathrm{d}}) - W(X_{\mathrm{c}})$, and unbiased spread $S(X) = [M(M-1)]^{-1} \sum_{m \neq n} d_1(x^{(m)}, x^{(n)})$, we report

$$\widetilde{D}_{\mathrm{mean}} = \frac{2\, d_2(\bar{x}_{\mathrm{d}}, \bar{x}_{\mathrm{c}})}{d_2(\bar{x}_{\mathrm{d}}, y) + d_2(\bar{x}_{\mathrm{c}}, y)}, \qquad \widetilde{D}_{\mathrm{dist}} = \frac{2D_{\mathrm{E}}}{S(X_{\mathrm{d}}) + S(X_{\mathrm{c}})},$$
$$R_{\mathrm{spread}} = \frac{S(X_{\mathrm{c}})}{S(X_{\mathrm{d}})}, \qquad R_{\mathrm{CRPS}} = \frac{\mathrm{CRPS}(X_{\mathrm{c}}, y)}{\mathrm{CRPS}(X_{\mathrm{d}}, y)}. \tag{5}$$

Table 7: **Inference cost of 50-member ensembles on one NVIDIA H100.** Marchuk-S2S and LaDCast generate 192 six-hourly states (48 days); FuXi-S2S is run at its native daily resolution to 42 days. Output steps are physical forecast times, not flow-integration steps; Marchuk-S2S uses 10 network evaluations per four-day chunk, 120 per 48-day member.

Model	Parameters	Time (min)	Output steps	Time per output step (s)
LaDCast	1.6B	73.3	192 (6-hourly)	22.9
FuXi-S2S	1.04B	5.1	42 (daily)	7.3
Marchuk-S2S	276M	9.7	192 (6-hourly)	3.0

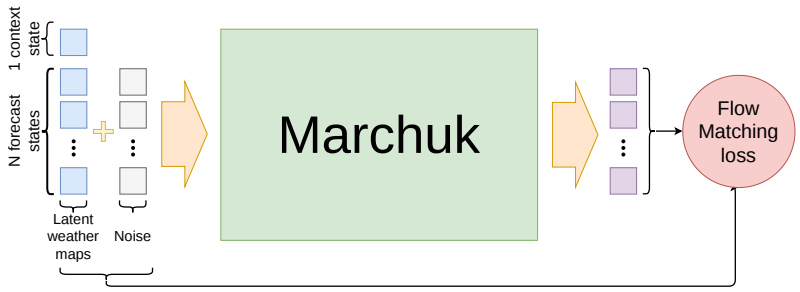

Figure 4: **Marchuk-S2S training and sampling.** One latent atmospheric state and the time-of-year labels of the source and target times condition a flow from an independent Gaussian sample to a joint trajectory of L future latent states. Resampling the noise produces ensemble members. At inference, the endpoint of each chunk conditions the next call.

The two discrepancies normalize differences in ensemble means and in full distributions by forecast error and ensemble width, respectively; identical ensembles give $(0, 0, 1, 1)$. For two independent ensembles from the same law, W includes the M zero diagonal terms, so $\mathbb{E}\, D_{\mathrm{E}} \approx 2S/M$ and $\widetilde{D}_{\mathrm{dist}} \approx 2/M$; the mean difference has variance $2\sigma^2/M$ against an error variance $\sigma^2(1 + 1/M)$ for a calibrated ensemble, so $\widetilde{D}_{\mathrm{mean}} \approx \sqrt{2/(M+1)}$. Metrics are computed per field and initialization, averaged arithmetically for discrepancies and in log space for ratios, and reported with standard errors over initializations.

A.7 CRPS Spread–Skill Ratio

Following the WeatherBench-2 CRPS implementation, the spread–skill ratio shown in the appendix figures is

$$\mathrm{SSR} = \frac{\mathbb{E}\|X - X'\|_{1,w}}{\mathbb{E}\|X - Y\|_{1,w}}, \tag{6}$$

where X and X' are independent members of the same ensemble, Y is the verifying ERA5 field, and $\|\cdot\|_{1,w}$ is the area-weighted L_1 norm. Values near one indicate that ensemble dispersion matches forecast error; values below one indicate underdispersion.

A.8 MJO Index

We evaluate MJO forecast skill with a modified RMM-like index, adapted from the Real-time Multivariate MJO (RMM) index (Wheeler & Hendon, 2004). The original RMM index combines the zonal winds U_{850} and U_{200} with outgoing longwave radiation (OLR) as a measure of tropical convection. OLR is neither an input nor an output of Marchuk-S2S, so we use 24-hour total precipitation as the convection proxy instead. For consistency, the same TP-based index is applied to the ensemble-mean forecasts of Marchuk-S2S, FuXi-S2S, and ECMWF and to the corresponding ERA5 fields, with forecasts aggregated to daily resolution as in Section 4.1. Skill is measured by the bivariate correlation (COR) between the forecast and ERA5 index time series (Lin et al., 2008), with $\mathrm{COR} = 0.5$ as the conventional skill threshold. Results are averaged over the 2019–2021 initializa-

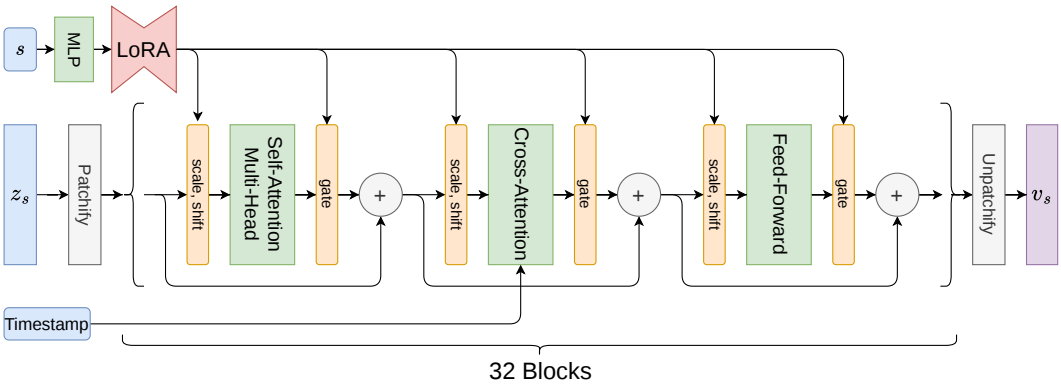

Figure 5: **Marchuk-S2S forecasting backbone.** Flow time controls the shared AdaLN–LoRA modulation, time-of-year embeddings enter through cross-attention, and 32 DiT blocks map the noisy latent trajectory to the conditional velocity.

tion dates. For all three systems, COR decreases with lead time and falls below 0.5 at approximately 30 days (Figure 3).

A.9 Additional Medium-Range Results

Figures 6–8 report the complete 15-day WeatherBench-2 evaluation on 732 initializations in 2020. Marchuk-S2S and LaDCast-1.6B have similar spread–skill ratios, both slightly below IFS-ENS at short leads.

A.10 Additional Subseasonal Results

Figure 10 shows the spread–skill ratio for the matched 2019–2021 subseasonal comparison of Figure 1. Marchuk-S2S is slightly underdispersive, with SSR between about 0.93 and 0.99, whereas the ECMWF S2S ensemble stays close to or slightly above one.

A.11 Case Studies

Figure 11 gives two qualitative views of forecast evolution and ensemble diversity. For three seasonally separated 2018 initializations, the ensemble follows the broad ERA5 seasonal evolution of extratropical area-mean 2-m temperature in both hemispheres over 48 days. Offsets from ERA5 of up to about 1 K are already present at the first lead time, so they do not arise from the rollout; this panel illustrates seasonal evolution rather than forecast error. For Hurricane Laura, a 50-member forecast initialized on 23 August 2020, four days before landfall, spans a range of tracks. A minority of members, including the closest member, reproduce the observed northward turn, while the ensemble-mean track is biased westward. These examples complement, but do not replace, the aggregate verification in the main text.

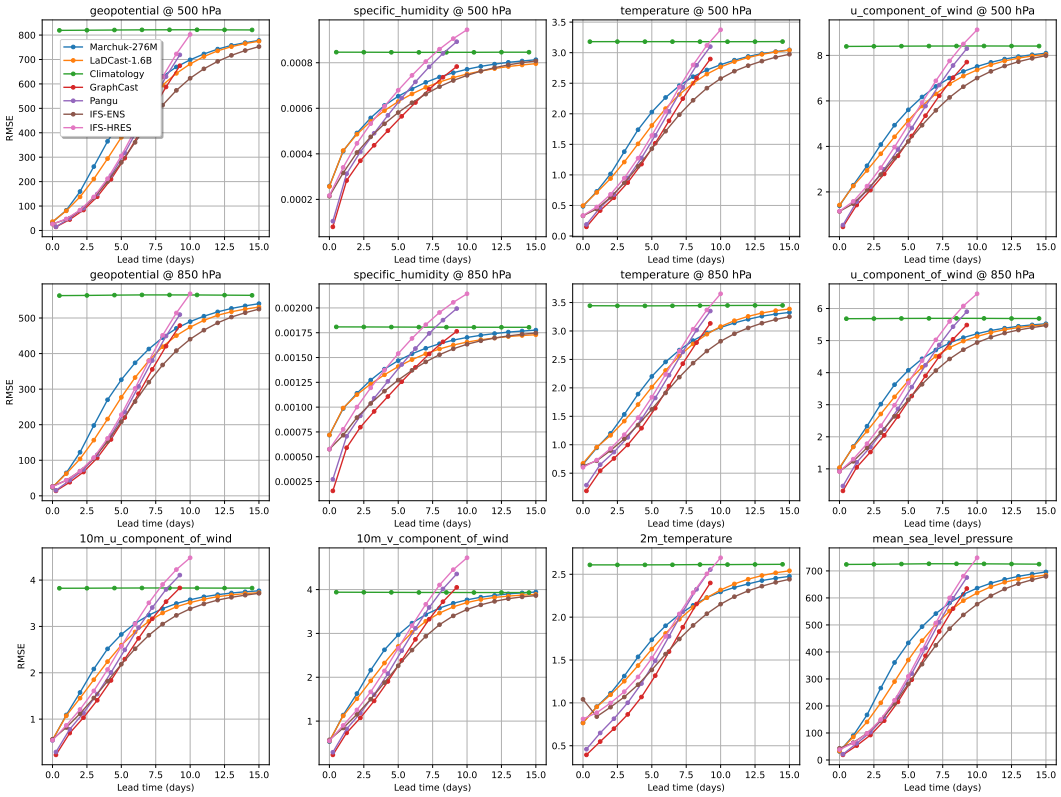

Figure 6: **15-day RMSE.** Latitude-weighted RMSE for Z_{500}, SH_{500}, T_{500}, U_{500}, Z_{850}, SH_{850}, T_{850}, U_{850}, 10m-UW, 10m-VW, T2M, and MSLP. Marchuk-S2S is compared with LaDCast-1.6B, GraphCast, Pangu-Weather, IFS-ENS, IFS-HRES, and climatology.

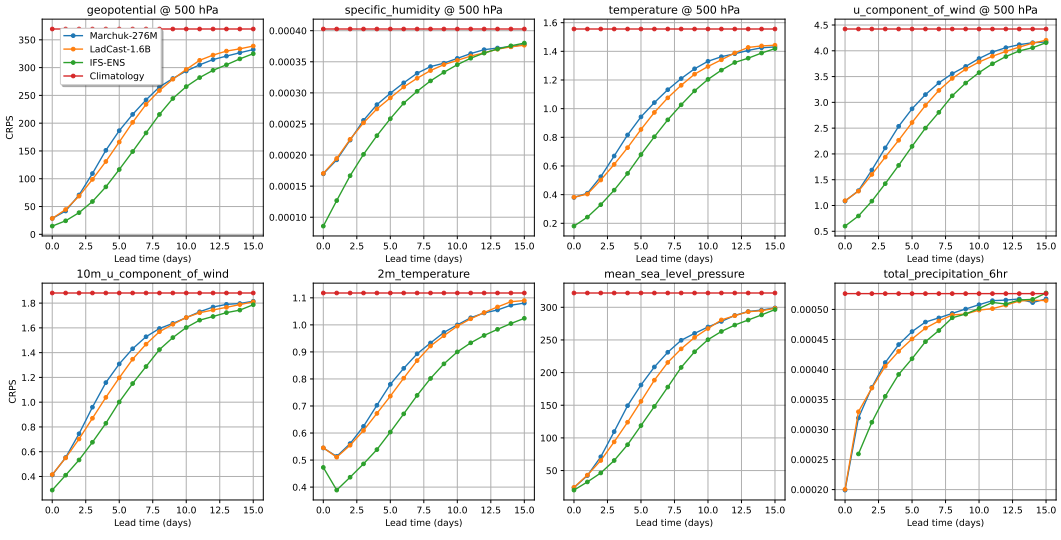

Figure 7: **15-day CRPS** for Z_{500}, SH_{500}, T_{500}, U_{500}, 10m-UW, T2M, MSLP, and 6-hour precipitation. Marchuk-S2S is compared with LaDCast-1.6B, IFS-ENS, and climatology.

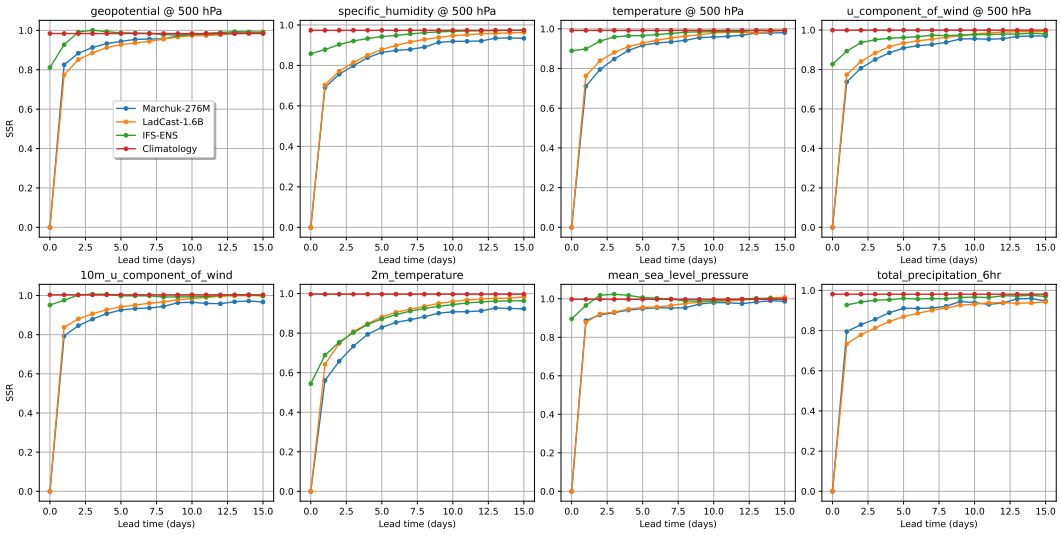

Figure 8: **15-day spread–skill ratio** as defined in equation 6, for the eight fields of Figure 7.

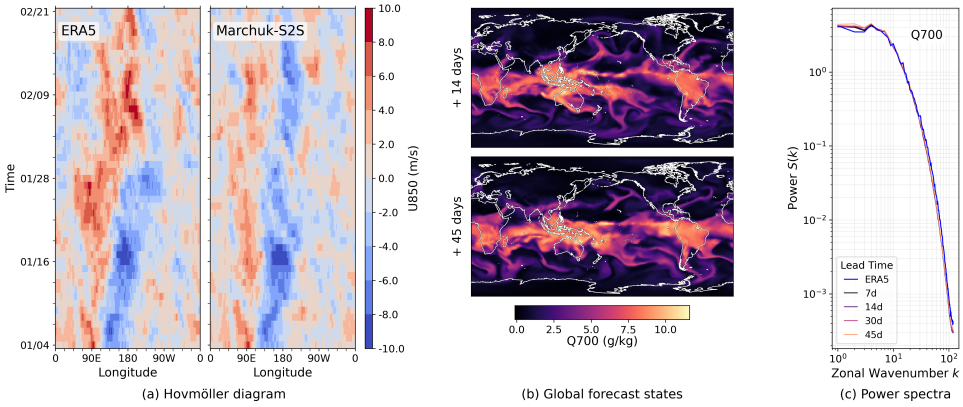

Figure 9: **Long-rollout structural diagnostics on 2018 forecasts.** (a) Hovmöller diagram (Hovmöller, 1949) of U_{850} anomalies averaged over 10°S–10°N for a forecast initialized on 4 January 2018. (b) Global Q_{700} fields of one member at days 14 and 45. (c) Zonal Q_{700} power spectra at days 7, 14, 30, and 45, averaged over 50 initializations and 50 members.

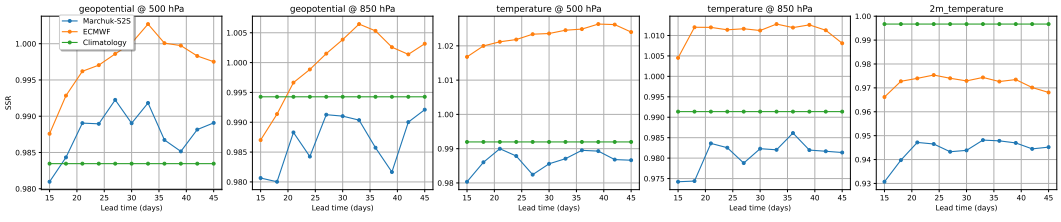

Figure 10: **Subseasonal spread–skill ratio** from days 15 to 45 on matched 2019–2021 initializations, for the fields of Figure 1.

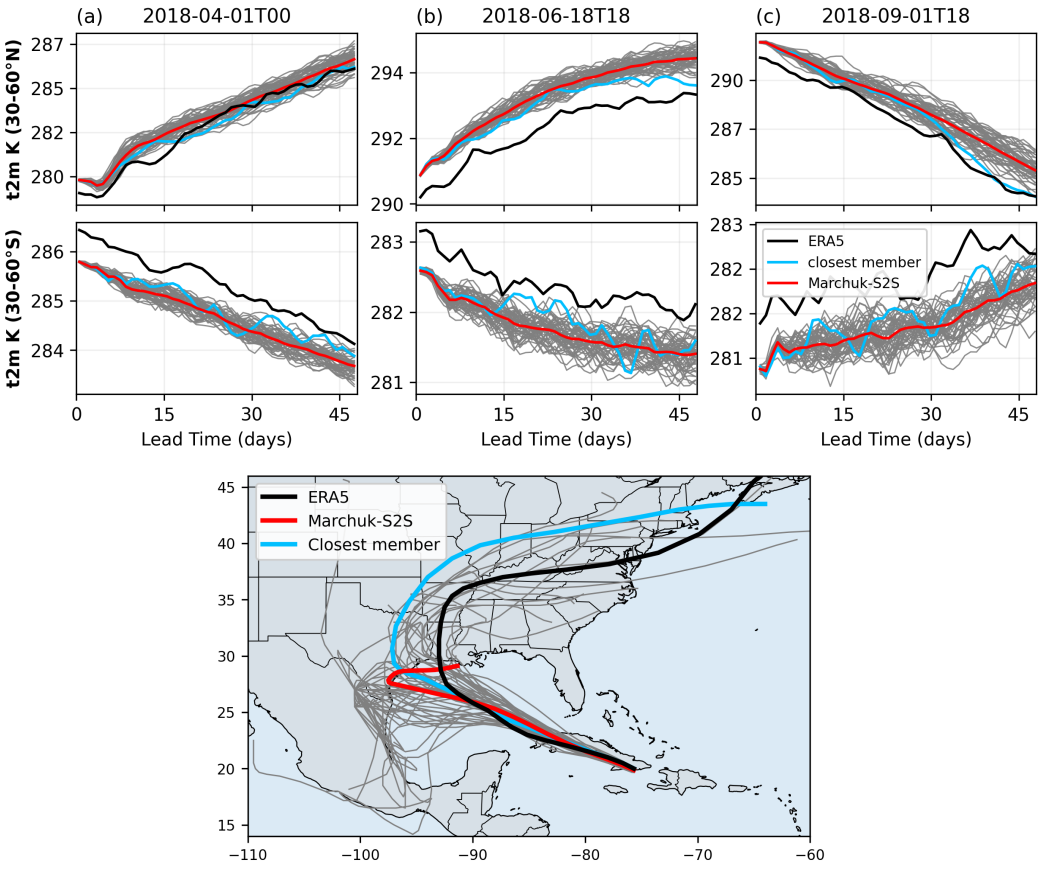

Figure 11: **Seasonal evolution and Hurricane Laura case study.** Top: area-mean 2-m temperature over the Northern (30°–60°N, upper row) and Southern (30°–60°S, lower row) Hemisphere extratropics for three 2018 initializations; gray curves are the 50 members, red the ensemble mean, cyan the closest member, and black ERA5. Offsets at the first lead are not forecast errors. Bottom: predicted tracks of Hurricane Laura from a 50-member forecast initialized on 23 August 2020; gray curves are members, black ERA5, red the ensemble-mean track, and cyan the closest member. Closest members are selected a posteriori against ERA5.