
Learning Ensemble Mixtures from Observations and Adapting Online with Theoretical Bounds

Saptarishi Dhanuka

Ashoka University, India

saptarishi.dhanuka@ashoka.edu.in

Sarvesh Iyer

Ashoka University, India

sarvesh.iyer@ashoka.edu.in

Manmeet Singh

Western Kentucky University

manmeet.singh@wku.edu

Mrutyunjay Mohapatra

Indian Meteorological Department

m.mohapatra@imd.gov.in

Parthasarathi Mukhopadhyay

Ashoka University, India

parthasarathi.mukhopadhyay@ashoka.edu.in

Sandeep Juneja

Ashoka University, India

sandeep.juneja@ashoka.edu.in

Abstract

Neural weather models have comparable aggregate performance to numerical weather predictors, but with heterogeneity in performance since no single model dominates benchmarks. This naturally motivates combining models for improvements, but existing methods are usually simplistic averages on coarse grids with little online adaptation and no formal error bounds. We learn the best model based on historical gridded forecast and ground *station* data, and adapt online with a learnt model based on recent errors, and an expert advice-based algorithm which has novel bounds for the final forecast error relative to the best-in-hindsight. Our best model gets 30+ % improvements in key metrics relative to standard methods. This inexpensive framework can be extended to any point observation set to create bespoke predictions customized to different use cases with formal guarantees, vital for climate change adaptation particularly in data sparse and compute constrained areas. The paper website and other materials can be found [here](#).

1 Introduction & Context

Neural weather models like FGN [1], FourCastNet [18], GenCast [19] and others have become compute-efficient competitors to traditional numerical weather predictors (NWP) on various metrics, and find use for critical applications in fields including agriculture, energy, and disaster management. Yet despite diverse architectural inductive biases, they show similar aggregate performance with no one model outperforming others consistently. The natural approaches to combine them include offline methods, which learn a combiner based on historical forecast-truth data, and online algorithms, which correct forecasts' outputs at every step based on recent errors. These are data driven by using model error patterns but frozen at deployment and offer no formal *guarantees* against distribution shift.

Online algorithms mitigate this by reweighing base predictors sequentially as observations arrive [25, 15, 10, 17]. NWP statistical post-processing methods like ensemble model output statistics (EMOS) [12], Bayesian model averaging (BMA) [20], and proper-scoring-rule calibration [11] recalibrate or combine existing forecast outputs using forecast-observation data,

making them simpler than learned supervised aggregators but less theoretically guaranteed than "prediction with expert advice" methods. Here, procedures for combining models give provably optimal results [23, 22, 9, 16, 4]. These are useful when analysing mixture errors against the best possible method in hindsight, defined as its *regret*, which leads to theoretically guaranteed bounds on quantities of interest in important applications [5].

Our extensible framework fills this gap by learning the best forecast from historical observations, then cheaply adapts online to data as it comes along. We develop two online adaptation methods, one having a novel formal bound developed by us chiefly in [6], that gives a guarantee on the error of the final forecast relative to the best forecast or possible forecast combination in hindsight, and the second which is a learnt recent window calibration. Both have better performance than a standard baseline used, with the learnt model empirically outperforming the former. We also do first evaluation of neural models on the Automated Weather Stations (AWS) network of the Indian Meteorological Department (IMD). Our best model observes **40 %** Continuous Ranked Probability Score (CRPS) improvements are observed in India, with more than **50 %** for challenging terrains in the hills, and even around 15-25 % for relatively more assimilated and data-rich regions like Europe and Continental United States (CONUS). We first describe our data, methods and experiments in Section 2. After stating our results in 3, we conclude in Section 4.

Pathway to Climate Impact: Many decisions are made locally at stations by stakeholders that cannot adapt a foundation weather model themselves, while models are constrained by the observations that are assimilated into their initial conditions or encoded into their weights through training. In sensitive areas where adaptive predictions are needed, our framework can be used by entities holding local data like renewable operators, disaster management, ports, etc. to obtain forecasts adapted to their site and use case without having to send data, retrain models, or operate heavy infrastructure, thus being important for climate change adaptation especially in challenging regions.

2 Data, Methods and Experiments

Forecast Data: 2019-2025 range (post ML models' training period) is taken from FourCastNet-3, GenCast AI weather models, and IFS-ENS, GEM, GEFS Numerical Weather Predictors, 2022-2025 from FGN, for 3-day lead times 00:00 UTC init times, 12 hour temporal resolution and 0.25° spatial resolution, over India, Europe and CONUS, for temperature and total accumulated precipitation variables, as available.

Station Data: India period 2019-2025 is from the IMD AWS network, while American and European data 2022-2025 is from Stationbench. More details in Appendix A.

Setup: For each forecast issue time $t \in [T]$, lead time $\tau \in \mathcal{T} = \{12, \dots, 72\}$ hours and station $s \in \mathcal{S}$, denote by $y_{t,\tau,s} \in [a, b] \subset \mathbb{R}$ the station's report at time $t + \tau$. There are N experts/models each carrying an atmospheric initial state x_t through a model pass $f_n : x_t \mapsto F_{n,t,\tau}$ giving a predictive CDF over a *grid*, and we use its crop over the station.

Offline Supervised Learning: We treat the forecast systems as a set so FGN can be added later from 2022. The model is a small set-transformer over per-system tokens with around 500,000 parameters, emitting 13 quantiles and trained with 15% system dropout. See Appendix B for more details.

Online Adaptation: We use two approaches, one is **VT-MOE**, a Vyugin-Trunov aggregator, which we motivate in Appendix B and chiefly developed in [6]. Second is a learnt combiner that takes the offline model as an expert alongside the raw systems, equal-weight, station climatology and Ensemble Model Output Statistics (EMOS). (See B for details). For a set A_r of experts, it keeps three parameters per key (a key being a (station, lead) pair), convex weights $w_i \geq 0, i = 1, \dots, |A_r|$ satisfying $\sum_{i=1}^{|A_r|} w_i = 1$, an additive shift b and a log-scale u . It pools the experts' quantile functions level by level, and then recalibrates about the pooled median m_r . While forecast belief is encoded by the weights w_i , the pair (b, u) corrects forecast sharpness at stations. The weights are updated by exponentiated gradient descent, and the affine pair by an AdaGrad step (see Appendix B) on the same loss.

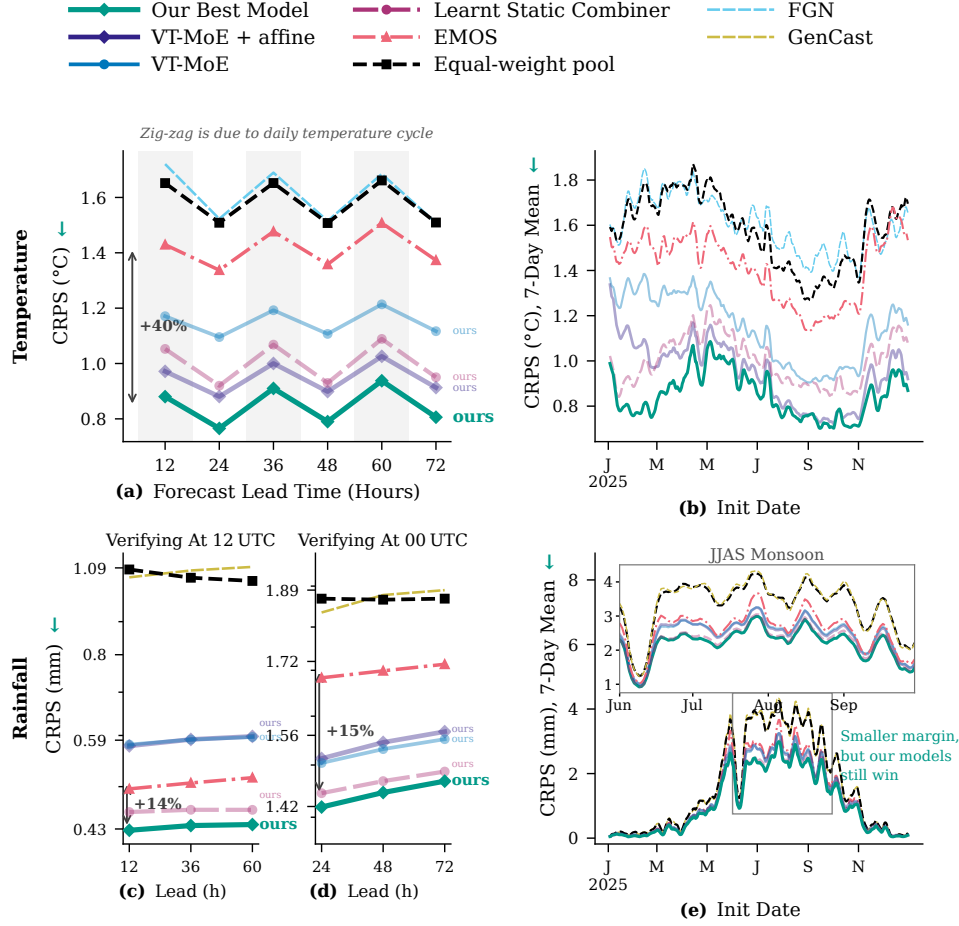


Figure 1: **Temperature, rainfall India CRPS:** (a, c) CRPS avg per forecast lead time; (b, e) the daily 7-day mean rollout shown for a 2025 subset of the eval period. The zig-zag in (a) due to the leads 12/36/60 h verifying at 12 UTC (shaded) and 24/48/72 h at 00 UTC

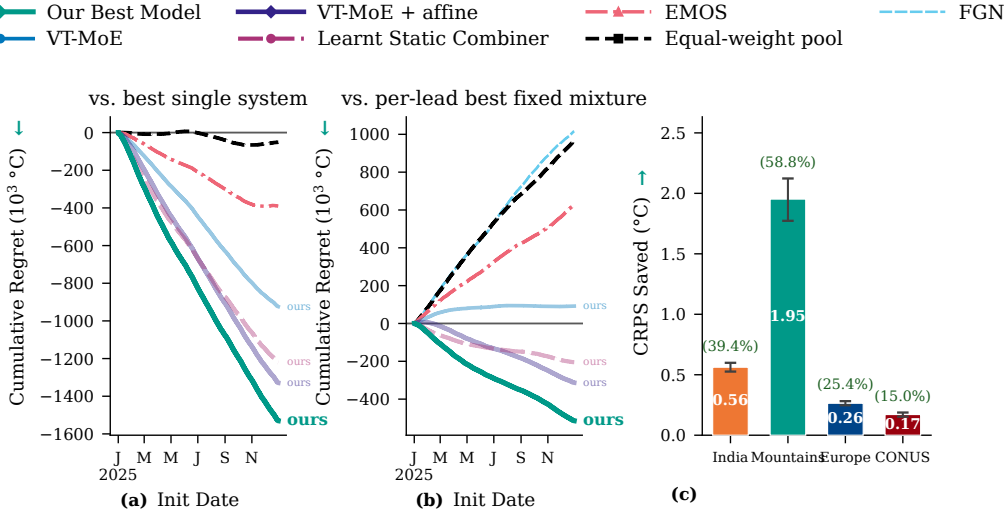


Figure 2: **Regret:** (a)(b) for temperature, learner adds affine correction so lies outside comparator class hence giving negative regret; (c) shows regionwise variation in saved temperature CRPS relative to EMOS

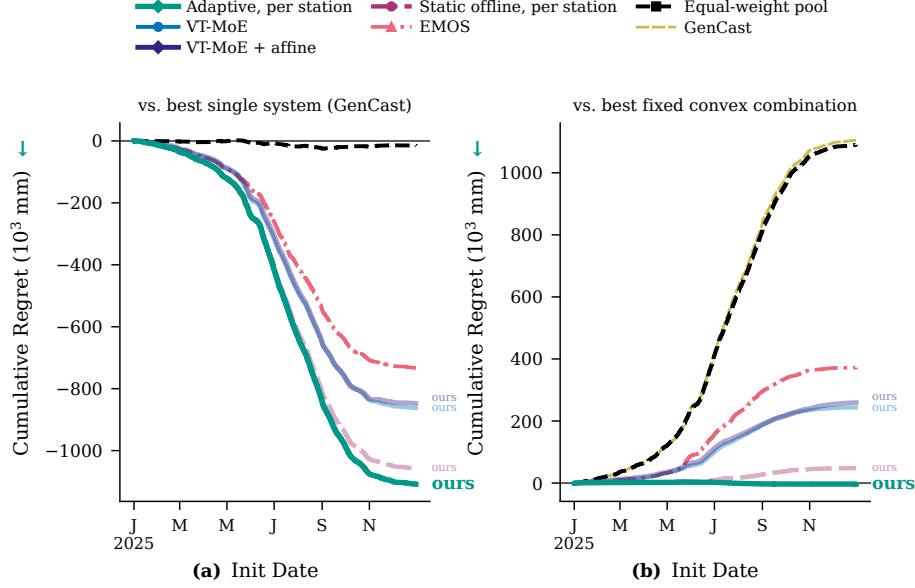


Figure 3: **Cumulative regret on rainfall**, against (a) the best single system and (b) the best fixed convex combination, both chosen in hindsight over the same window. Below zero beats a comparator no forecaster could have issued.

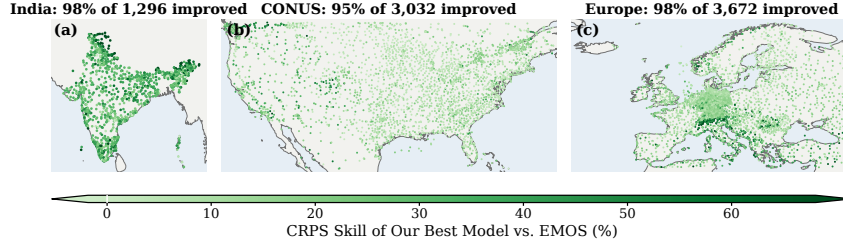


Figure 4: **Improvement over standard EMOS method**

Feedback is delayed, in that a lead- L forecast verifies only L hours later. This is referred to as "Our Best Model" in the plots.

Experiments and Evaluation: The offline model is trained from 2019-2023 forecast-ground truth pairs for India (2022-2023 for the US and Europe), and evaluated on 2024-2025, where we run an online adaptation algorithm. All forecasts are evaluated primarily on the CRPS. During evaluation, a 3 day forecast is initialised everyday and then measured against the matched station observations. See Appendix B for complete details.

3 Results

Figure 1 shows our models having lower CRPS than other methods for both temperature and precipitation, with the improvement being relatively larger for temperature due to the challenging sparse nature of precipitation. We are considerably better than equal-weight and standard EMOS bias correction methods. Our provable algorithm, while not the best experimental line on the graph, still beats the two aforementioned baselines, with the added benefit of the theoretically guaranteed regret bound. In (b), (e) we also see the error rolled out over one of the test years, with the largest gains for precipitation coming during the core monsoon months of June, July, August and September.

Figure 2 (a) (b) show our method having lower regret than others for temperature, with similar results for precipitation shown. Fig 2 (c) shows improvement across India, Europe and CONUS, with mountainous regions contributing a higher relative performance gain. We

show that our improvements are not just cause of simple terrain elevation bias correction through ablations in the appendix. Figure 4 visually shows improvements relative to other models, with significant improvements visible in North India and central Europe. These could also signify data attribution differences depending on location and elevation, in the reanalysis data these models were trained on, to explain the differences in performance.

Theoretical Bound. Let $\Theta = \{(w_1, \dots, w_N) : w_i \geq 0, \sum_{i=1}^N w_i = 1\}$ be simplex of fixed convex mixtures of N experts. For fixed time $t \in [T]$, assume that i th expert CDF is given by $F_{i,t}$. For $p = (p_1, \dots, p_N) \in \Theta$ define the mixture CDF $F^p(t) = \sum_{i=1}^N p_i F_{i,t}$. The *regret* for VT-MOE over best mixture in hindsight is defined by

$$\text{Reg}_T = \sum_{t \leq T} \text{CRPS}(F_t, y_t) - \min_{p \in \Theta} \sum_{t \leq T} \text{CRPS}(F_t^{(p)}, y_t).$$

In Appendix E, we prove for observations coming from $[a, b]$, assuming that VT-MOE can integrate exactly over Θ , that,

$$\text{Reg}_T \leq \frac{1}{2}(b-a)(N-1) \ln T + C, \text{ where } C = C(a, b, N) \text{ is independent of } T. \quad (1)$$

4 Conclusion

We show a framework that learns a forecast from historical forecast-observation pairs, followed by online adaptation with both a learnt adapter, and an algorithm that has formal relative error bounds on the combination method, outperforms other online bias correction methods on key metrics. This can be extended to any station data with appropriate data matching and geocoded metadata, with extensions to other datasets planned. Due to higher improvements observed in mountains with complex terrain, we also plan to use land use data and earth observation embeddings like AlphaEarth [2] and Tessera [8], to enable a more unified representation of Earth by linking atmospheric models with the ground and the surrounding context, further improving predictions at point values like stations, renewable locations, and weather sensitive infrastructure in a stakeholder oriented manner.

References

- [1] 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.
- [2] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.
- [3] Nicolás Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, may 1997.
- [4] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [5] Kai Chang and Themistoklis P Sapsis. Extreme event aware (η -) learning. *Nature Communications*, 2026.
- [6] Saptarishi Dhanuka, Sarvesh Iyer, Manmeet Singh, Mihir More, Rushil Gupta, Dhruvan Gupta, Parthasarathi Mukhopadhyay, and Sandeep Juneja. Adaweather: Adaptively mixing probabilistic weather forecasts with logarithmic regret, 2026.
- [7] Clément Dombry and Ahmed Zaoui. Distributional regression: CRPS-error bounds for model fitting, model selection and convex aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [8] Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 34818–34831, 2026.
- [9] Yoav Freund. Predicting a binary sequence almost as well as the optimal biased coin. Technical report, AT&T Research, 1996.
- [10] Yuwei Fu, Di Wu, and Benoit Boulet. Reinforcement learning based dynamic model combination for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
- [11] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- [12] Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.
- [13] David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Tight worst-case loss bounds for predicting with expert advice. *Technical Report UCSC-CRL-94-36, University of California, Santa Cruz*, 1994.
- [14] Elad Hazan. *Introduction to Online Convex Optimization*. Foundations and Trends in Optimization, 2016.
- [15] Ying-ye Ava Lau, Zhiwen Shao, and Dit-Yan Yeung. Fast and slow streams for online time series forecasting without information leakage. In *International Conference on Learning Representations (ICLR)*, 2025.
- [16] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

- [17] Larkin Liu and Jalal Etesami. Online mixture of experts: No-regret learning for optimal collective decision-making. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [18] 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.
- [19] 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. GenCast: Diffusion-based ensemble forecasting for medium-range weather. *Nature*, 2024.
- [20] Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.
- [21] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
- [22] Vladimir G. Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 51–60. ACM, 1995.
- [23] Volodimir G Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–386. Morgan Kaufmann, 1990.
- [24] Vladimir V. V'yugin and Vladimir Trunov. Online aggregation of probability forecasts with confidence. *arXiv preprint arXiv:2109.14309*, 2021.
- [25] Yi-Fan Zhang, Qingsong Wen, Xue Wang, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. OneNet: Enhancing time series forecasting models under concept drift by online ensembling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Appendix

A Datasets

In the below table, we organize the data utilized in this article by their name, developer and time period.

Table 1: **Forecast and station datasets used.** All forecasts are initialised daily at 00 : 00 UTC and are at 0.25° resolution.

Name / Domain	Developer / Network	Time period	Stations
<i>Forecast experts</i>			
FourCastNet-3	NVIDIA	2019–2025	N.A.
FGN	Google Research	2022–2025	N.A.
GenCast	Google DeepMind	2019–2025	N.A.
IFS ENS	ECMWF	2019–2025	N.A.
GEM	ECCC / MSC	2019–2025	N.A.
GEFS	NOAA NCEP	2019–2025	N.A.
<i>Station observations</i>			
Domain	Network	Time period	Stations
India	IMD AWS	2019–2025	1296
Europe	Stationbench	2022–2025	3672
CONUS	Stationbench	2022–2025	3032

AWS means Automated Weather Station Network of the IMD. This is also the first analysis of the AI weather models on this station data. Stationbench is an open-source Python library for benchmarking weather forecasts against weather station data.

Ensemble sizes are the member count over the windows used. Two of them differ between the crops of the archive we hold: GenCast is 52 members over India and 48 over Europe and CONUS, and FourCastNet-3 is 15 and 16 respectively. The station counts are the ones that survive the filter (at least 200 verifications in the window)

The domains are the boxes the forecast crops are cut to, and the station sets follow the crops rather than the other way round: India $6\text{--}38^\circ\text{N}$, $68\text{--}98^\circ\text{E}$; Europe $35\text{--}72^\circ\text{N}$, $25^\circ\text{W}\text{--}45^\circ\text{E}$; CONUS $24\text{--}50^\circ\text{N}$, $125\text{--}66.5^\circ\text{W}$. Temperature is taken from all the models, and precipitation from Gencast, GEFS, and GEM.

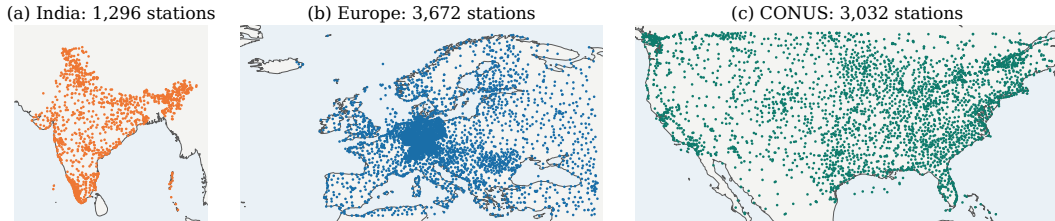


Figure 5: **The three verification networks.**

Why combine at all. The premise of the whole framework is that no single system is reliably best. Figure 6 resolves that station by station: on temperature FGN wins at 44 % of Indian stations and GEM at 22 %, with the rest split three ways; on rainfall GEFS and GenCast split the country almost evenly.

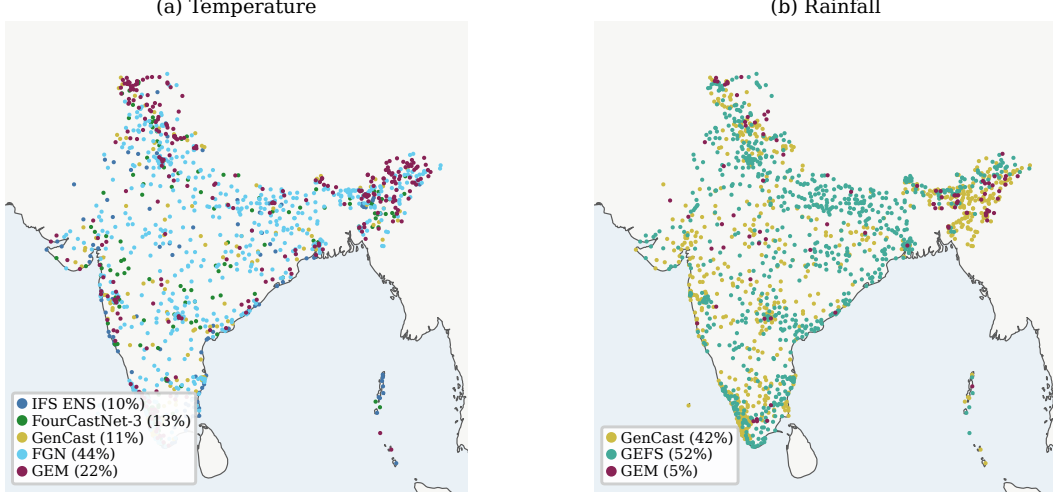


Figure 6: No single system is best everywhere.

B Methods

Setup. Index forecast issue times by $t \in [T]$, lead times by $\tau \in \mathcal{T} = \{12, \dots, 72\}$ h and stations by $s \in \mathcal{S}$. For each triple there is a scalar observation $y_{t,\tau,s} \in [a, b] \subset \mathbb{R}$, the station’s report at valid time $t + \tau$. Every method emits a predictive quantile function on a shared grid of 13 levels, $\mathcal{A} = \{0.025, 0.05, 0.1, \dots, 0.9, 0.95, 0.975\}$, obtained for a raw ensemble by interpolating each member to the station point and taking quantiles across members *at the station*. Taking grid-point quantiles first would narrow every predictive distribution before it reached the station.

Offline Supervised Combiner A two-layer pre-norm transformer ($d = 128$, four heads) attends over the tokens and the encoded tokens are mean-pooled over the set, concatenated with a row context (the anchor, the dispersion of the experts’ means, how many are present, the lead, day-of-year and hour harmonics, position and elevation), and passed to a head emitting the 13 quantile levels. Model is around 500K parameters, trained on 2019-2023 for India, 2022-2023 for others, with 15% of the available systems hidden at random per batch for robustness to missing forecasts.

A mask over systems enters the encoder, so a system that does not exist on a given day is absent rather than imputed, and FGN can join the panel in 2022 without retraining. And the head emits cumulative positive increments, $q(\alpha_{k+1}) = q(\alpha_k) + \text{softplus}(\cdot)$, so quantile crossing is impossible by construction rather than by penalty. The output is written as anchor + offsets, where the anchor is the equal-weight mean of the awake systems’ means, so what the network learns is explicitly a *correction* to a naive pool.

Online Adaptation The offline combiner is handed to the online layer as one expert among N , alongside the raw systems, their equal-weight pool, a causal expanding station climatology and EMOS. For a row r with awake expert set A_r the layer holds three parameters per key, a key being a (station, lead) pair: convex weights $w \in \Delta^{|A_r|}$, an additive shift b , and a log-scale u . It pools the experts level by level (Vincentization) and recalibrates about the pooled median $m_r = \bar{q}_r(0.5)$:

$$\bar{q}_r(\alpha) = \sum_{k \in A_r} w_k q_{r,k}(\alpha), \quad \tilde{q}_r(\alpha) = m_r + e^u (\bar{q}_r(\alpha) - m_r) + b.$$

The pooled quantile function is linear in w and the pinball loss is convex, so the weights take an exponentiated-gradient step and (b, u) an AdaGrad step on the same loss. The EG rate is the adaptive one of [4], $\eta = \min\{1/L_{\max}, 1.0739\sqrt{\ln N/V}\}$ with V the accumulated weighted variance of the expert losses, so there is no step size to tune — and therefore no opportunity to tune one on the evaluation window.

We remark briefly on the algorithm AdaGrad. It is an per-coordinate gradient descent with learning rate scaled by the inverse square root of accumulated squared gradients, which removes the need to tune separate rates for the shift and log-scale and anneals each key's step size as evidence accrues)

Feedback is *delayed*, and the delay is respected exactly: a lead- L forecast issued at t verifies only at $t + L$, so its loss may enter the state only once an initialisation has passed that valid time. The convenient alternative — letting a 72 h forecast be issued by weights that have already seen its own verification — is what makes an online result look better than it is, and is avoided throughout.

VT-MoE. More details in appendix E.

EMOS, the baseline that matters. Ensemble model output statistics [12] is the post-processing usually used operationally so we take it as a baseline. We fit the Gaussian form

$$y \sim \mathcal{N}(\mu, \sigma^2), \quad \mu = a + b\bar{m}, \quad \sigma = \exp(c + d \log(1 + \bar{s})),$$

where $\bar{m}$ is the consensus mean of the awake systems' ensemble means and $\bar{s}$ the mean of their spreads. Using the consensus rather than one coefficient per system is deliberate cause it keeps the baseline defined in every year, including those in which a system is absent, which a per-system parameterisation is not. The four coefficients are fitted by *direct CRPS minimisation* on the fitting window — not by maximum likelihood — using the closed form for a Gaussian,

$$\text{CRPS}(\mathcal{N}(\mu, \sigma^2), y) = \sigma \left[z(2\Phi(z) - 1) + 2\varphi(z) - \frac{1}{\sqrt{\pi}} \right], \quad z = \frac{y - \mu}{\sigma},$$

so the baseline is optimised for the metric it is judged by and cannot be dismissed as having been fitted for something else.

Evaluation metric and training loss. The quantile (pinball) loss at level α is

$$\rho_\alpha(y, q) = \max\{\alpha(y - q), (\alpha - 1)(y - q)\} = (\alpha - \mathbf{1}\{y < q\})(y - q), \quad (2)$$

and the CRPS is twice its integral over levels [11],

$$\text{CRPS}(q, y) = 2 \int_0^1 \rho_\alpha(y, q(\alpha)) d\alpha = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}\{z \geq y\})^2 dz, \quad F = q^{-1}. \quad (3)$$

Every fitted component in this paper, namely offline combiner, the EMOS coefficients, the online weights and the affine pair, all minimises (2) averaged over the $L = 13$ stored levels, which is a discretisation of (4): the training loss is the evaluation metric, so nothing is optimised for a proxy.

C Europe, USA Results

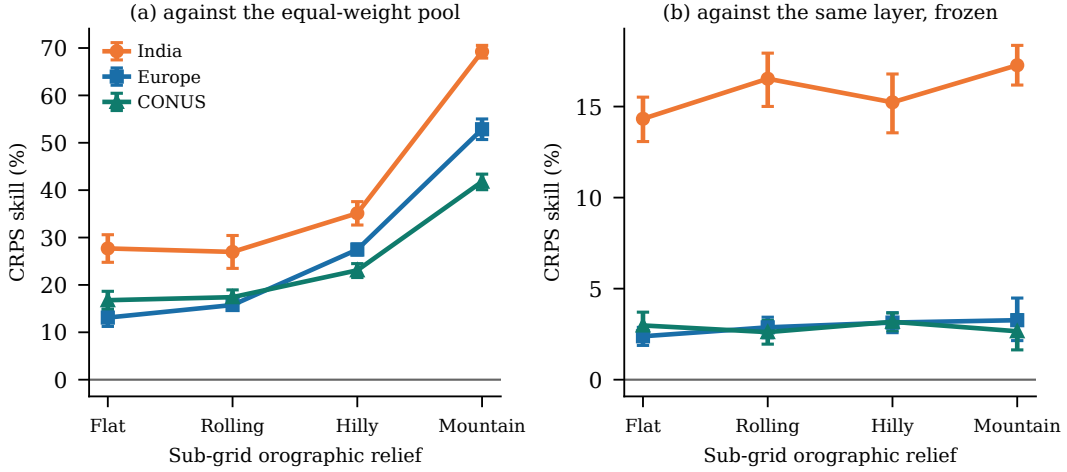
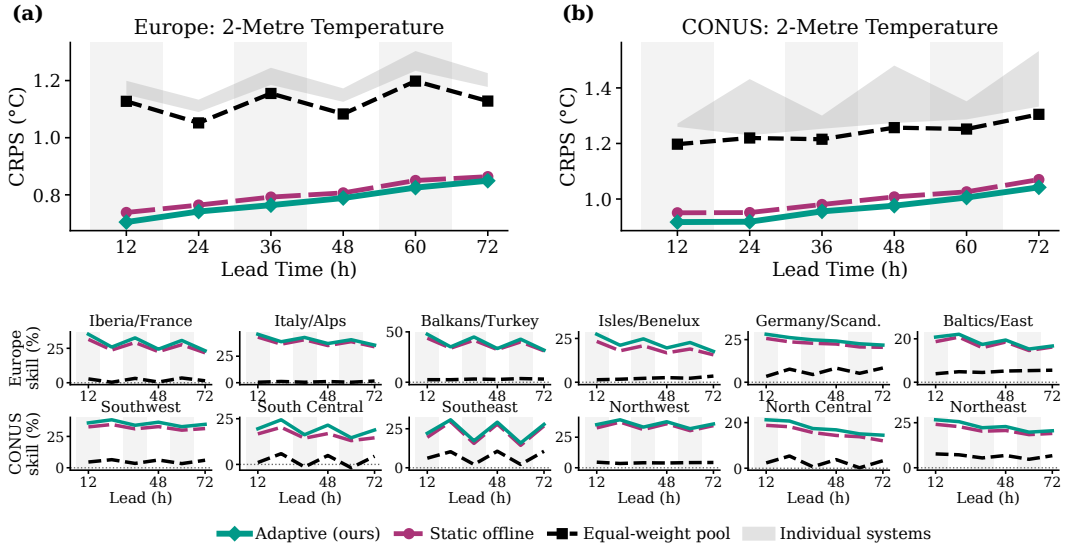


Figure 7: **Skill against sub-grid orographic relief, three networks.** Higher is better. (a) The gain over an equal-weight pool rises steeply with relief on every network (b) The gain from adapting online rather than freezing the same layer is far flatter, and several times larger in India which shows the difference between a network the global models already assimilate well and one they don't



Skill is against the reference chosen on each fitting window (Europe gencast, CONUS gem). Shaded leads verify at 12 UTC.

Figure 8: **The body's Figure 1, on two more continents.**

D Ablations

Which component actually goes stale. The layer holds convex weights w and an affine correction (b, u) . Freezing one while the other adapts says which one’s staleness the gain comes from.

Table 2: **Freezing one component at a time**, temperature CRPS ($^{\circ}\text{C}$) over India.

	Frozen correction	Adaptive correction
Frozen weights	1.037	0.906
Adaptive weights	0.965	0.872

Adapting only the affine correction recovers most of the available gain ($1.037 \rightarrow 0.906$) while adapting only the weights recovers markedly less ($1.037 \rightarrow 0.965$). What goes stale is chiefly where and how sharply a forecast sits at a station, not which forecast to believe — which is why the (b, u) pair is in the layer at all.

Why the affine term beats VT-MoE, and why VT-MoE is still useful The two online arms differ in exactly one respect: the hypothesis class. VT-MoE predicts inside Θ , the simplex of fixed mixtures; the deployed layer adds the affine pair (b, u) and therefore predicts outside it. The VTMOE algorithm underperforms the learnt online combiner, which is not a defect of the algorithm. No mixture of these experts, not even the best one chosen with full knowledge of the answer, can express a correction that differs station by station and a 0.25° grid box is wrong at a station in a way that is specific to that station. VT-MoE is optimal against the class it is defined over and carries the $O(\log T)$ guarantee for doing so, which the exponentiated-gradient layer does not; what it cannot do is leave that class. The practical reading is that the two are complementary: *the bound tells you how much you can lose by not knowing the best mixture in advance, and the affine term is what gets you past the best mixture.*

Where the gains comes from Figures 9 and 10 break the reduction down by lead, region, season and comparator simultaneously Rainfall is not uniformly positive and we do not claim it is.

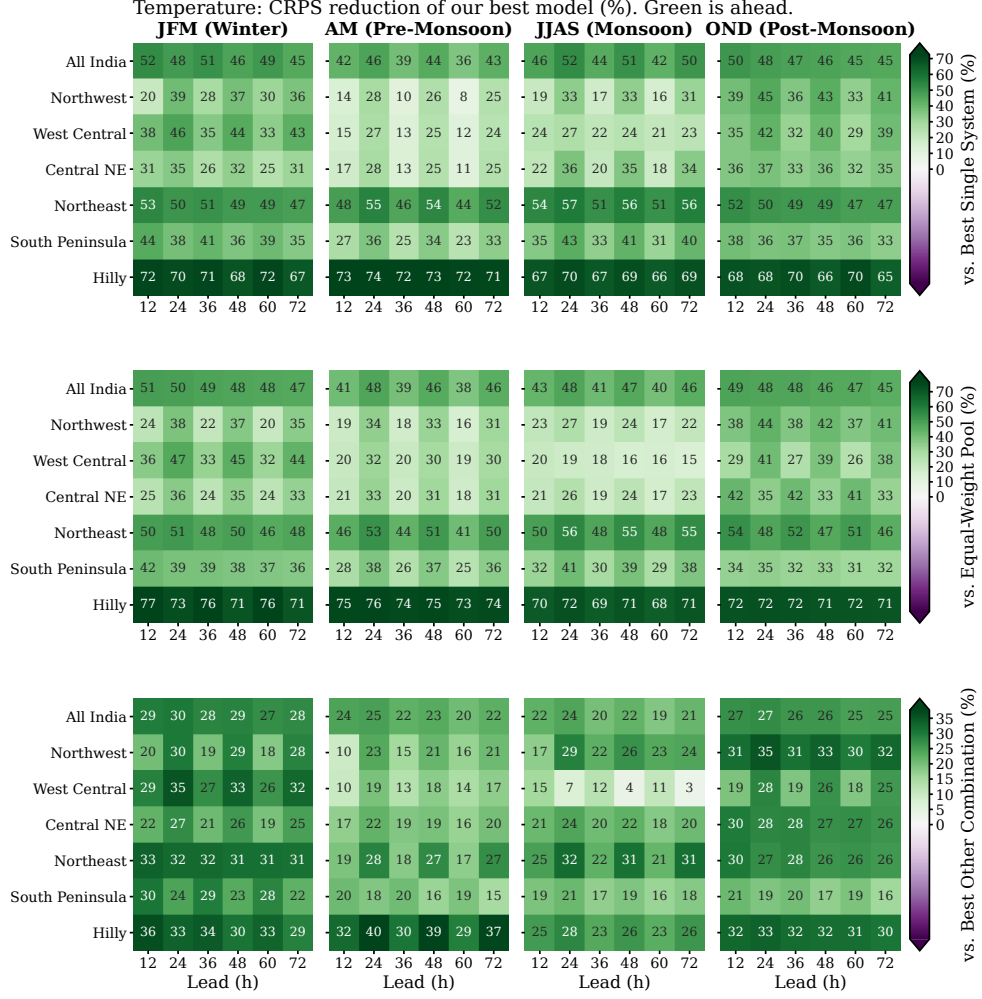


Figure 9: **Temperature: CRPS reduction of our best model relative to EMOS (%)**, by lead (columns within a block), IMD homogeneous region (rows) and season (blocks). Green is ahead.

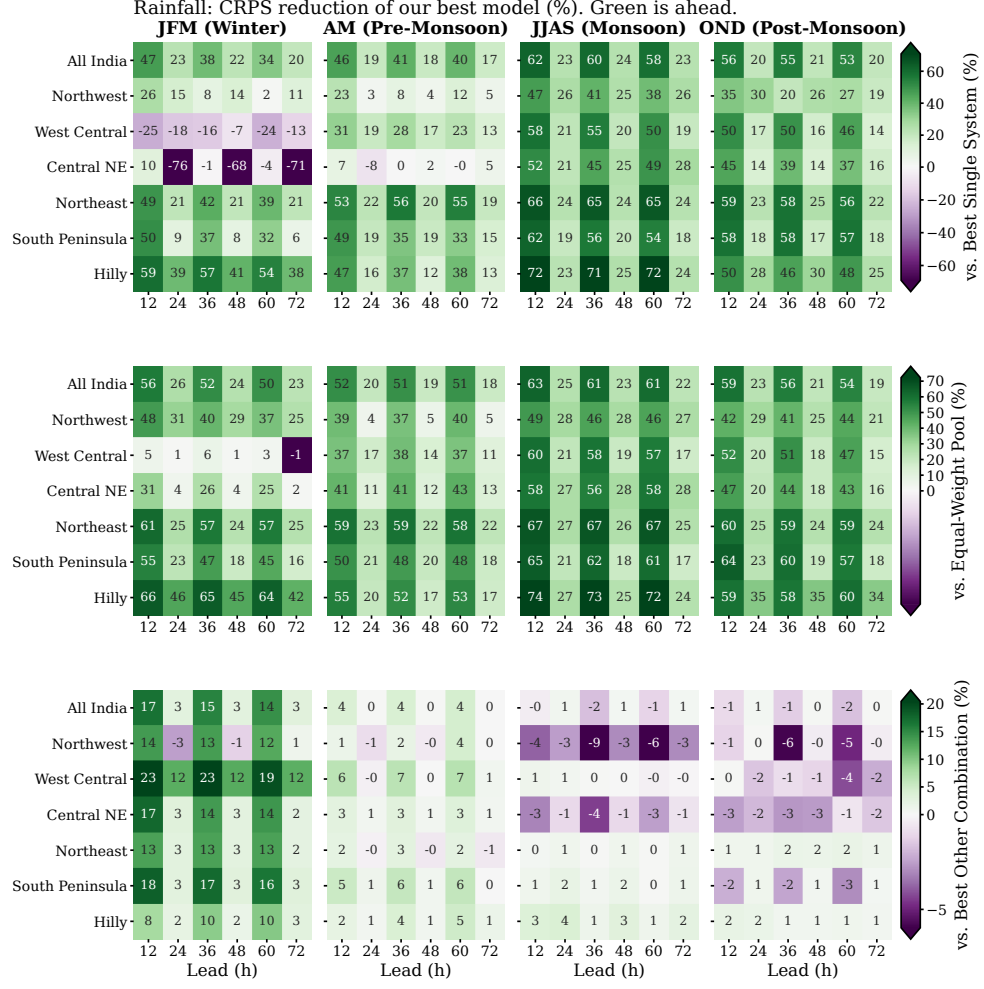


Figure 10: **Rainfall, same construction for India, relative to EMOS** The alternating light and dark columns are the two verification hours, 00 and 12 UTC. Purple cells are losses: unlike temperature, rainfall is not positive everywhere, and the exceptions sit in dry seasons in dry regions where the denominator is small.

Bringing your own data.

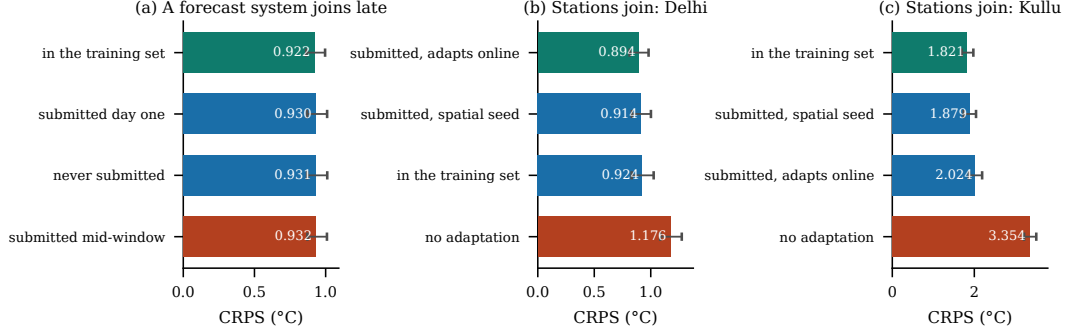


Figure 11: **What submitting data buys.** Lower is better (a) A forecast *system* (FGN) is withheld from the panel and then submitted part-way through. Scored on the rows after the submission date, arriving mid-window is worth as much as having been present from the first day, and both sit inside the interval of never submitting it at all: one expert among nine moves little, and the expert is not in the combiner’s training set in any submitted arm, so withholding it costs only what leaving one trained expert out costs. (b,c) A user’s own *stations* are submitted instead, at two contrasting sites, namely 74 stations near Delhi on the plain and 65 near Kullu in the western Himalaya. Here the picture is different: adapting online to them beats the frozen layer by a wide margin at both, and at Kullu seeding a new station from its neighbours (1.879 °C) recovers most of the gap between adapting cold (2.024) and having trained on it all along (1.821). At Delhi, adapting online is already better than having trained on the station.

E Formal Theoretical Math Analysis and Proofs

E.1 Preliminaries

Forecasts and observations. We work in the probabilistic forecasting setup with multiple forecast trajectories generated for each day by each model. As done previously, we index forecast issue times by $t \in [T]$ and lead times by $\tau \in \mathcal{T}$. For each pair (t, τ) , we observe a target field $y_{t,\tau} \in \mathcal{Y} \subseteq \mathbb{R}^{|\mathcal{G}|}$ on a spatial station grid $\mathcal{G}$, with each grid-point value lying in a known interval $[a, b] \subset \mathbb{R}$. Boundedness is realistic for variables such as temperature and is essential for the bounds we derive. There are N base forecasters, or *experts*. For each (t, τ) , expert n outputs a predictive CDF $F_{n,t,\tau}$ over $\mathcal{Y}$, derived from an atmospheric initial state x_t via a model pass $f_n : x_t \mapsto F_{n,t,\tau}$. In practice $F_{n,t,\tau}$ is represented by an M -member ensemble and we work with the empirical CDF; our analysis imposes no structure and requires no accuracy bound on expert predictions.

Loss. Quality is measured by the CRPS (CRPS) [11], evaluated pointwise and averaged over $\mathcal{G}$. For any predictive CDF $\hat{F}$ and observation y at a single grid point,

$$\text{CRPS}(\hat{F}, y) = \int_a^b (\hat{F}(z) - \mathbf{1}\{z \geq y\})^2 dz. \quad (4)$$

For an empirical CDF given by an ensemble $\{e_1, \dots, e_M\}$, CRPS admits the closed form $\frac{1}{M} \sum_i |e_i - y| - \frac{1}{2M^2} \sum_{i,j} |e_i - e_j|$ (this follows from arguments analogous to those in Lemma 4); in implementation we use the unbiased $M(M-1)$ -normalised plug-in estimator (details in Appendix E). We let CRPS act on grid-valued $(\hat{F}_{t,\tau}, y_{t,\tau})$ by averaging (4) over $g \in \mathcal{G}$, and write the cumulative loss as

$$\mathcal{L}_T = \sum_{t=1}^T \sum_{\tau \in \mathcal{T}} \text{CRPS}(\hat{F}_{t,\tau}, y_{t,\tau}).$$

CRPS is a strictly proper scoring rule [11] and is $\frac{2}{b-a}$ -mixable on $[a, b]$ [24, 7]. It also carries a quadratic structure (Lemma 4) that is central to controlling its fluctuations. Both properties drive the logarithmic regret bound of Section E.3.1.

E.2 Aggregation algorithms

An aggregation algorithm $\mathcal{A}$ produces a predictive CDF $\hat{F}_{t,\tau} = \mathcal{A}(x_t, \{F_{n,t,\tau}\}_{n=1}^N)$ from the initial state and the expert predictions; see [22],[24] for canonical examples.

Online aggregation. At each step the algorithm plays a CDF $\hat{F}_t$ and incurs loss $\text{CRPS}(\hat{F}_t, y_t)$, then updates once y_t is revealed. A canonical family of online aggregators is the *linear mixture* indexed by weights $w \in \Theta := \{w \in \mathbb{R}_{\geq 0}^N : \sum_n w_n = 1\}$, $\hat{F}_t^{(w)} = \sum_{n=1}^N w_n F_{n,t,\tau}$. Our key methodological contribution is a second online procedure producing a forecast that is non-linear in the expert predictions *and* the weights (Section E.3.1).

Regret. We measure online performance by the static regret against the best fixed convex combination in hindsight,

$$R_T = \sum_{t=1}^T \text{CRPS}(\hat{F}_t, y_t) - \min_{w^* \in \Theta} \sum_{t=1}^T \text{CRPS}(\hat{F}_t^{(w^*)}, y_t). \quad (5)$$

This is strictly stronger than regret against the best single expert, since Θ contains the vertices e_n . For best-expert regret on mixable losses, an algorithm utilising mixability achieves an optimal $O(\ln N)$ bound independent of T [22, 13, 9, 4]; for convex losses the rate degrades to $O(\sqrt{T \ln N})$ [21, 14]. In Section E.3.1 we derive an aggregator whose regret (5) grows $\ln(t)$ with an explicit leading constant; we discuss optimality of this constant in Section E.6 via comparison to the two-expert setting in [9].

E.3 Online adaptation

At inference time we treat the trained offline model as an additional expert and combine it online with the other N forecasters. We suppress the lead-time index τ throughout the analysis, but all results apply uniformly across lead times. We describe two online aggregators: Vovk’s algorithm as a baseline, covered in Appendix E.5 and our proposed VT-MOE. Vovk achieves $O(\ln N)$ regret against the best single expert [22, 13], but degrades against the strictly stronger best-mixture comparator (5): weights are maintained on individual experts and not on their mixtures, while the non-linearity of CRPS means a mixture can substantially outperform every individual expert (see Lemma 4). Lower bounds of $\Omega(\sqrt{T})$ are known for related losses [3]. This motivates an algorithm that tracks weights over mixtures directly.

E.3.1 VT-MOE: aggregation for mixtures of experts

Our proposed algorithm, Vyugin-Trunov Mixture of Experts (VT-MOE), adapts the Vyugin–Trunov aggregator [24] to the mixture-of-experts comparator. Write $F_t^{(p)} = \sum_{n=1}^N p_n F_{n,t}$ for the predictive CDF induced by mixture $p \in \Theta$. The mixability framework underlying the aggregator needs to be tailored to mixtures.

Definition 1. For $\eta > 0$, the CRPS loss is η -mixable in the MOE setting if, for every collection of expert CDFs $\{F_n\}_{n=1}^N$ and every finite measure ν on Θ , there exists a CDF F such that for all $y \in [a, b]$,

$$\text{CRPS}(F, y) \leq -\frac{1}{\eta} \ln \frac{\int_{\Theta} \exp(-\eta \text{CRPS}(F^{(p)}, y)) d\nu(p)}{\nu(\Theta)}. \quad (6)$$

VT-MOE maintains a measure μ_t on Θ , initialised to the uniform prior μ_0 (a choice supported by the two-expert analysis of 9, Sections 4,7). At round t it outputs a CDF $\hat{F}_t$ satisfying (6) relative to μ_t , then updates

$$\frac{d\mu_{t+1}}{d\mu_t}(p) = \exp(-\eta \text{CRPS}(F_t^{(p)}, y_t)). \quad (7)$$

Theorem 2 (Section E.4) establishes that such an $\hat{F}_t$ exists in closed form, with $\eta = 2/(b-a)$.

Closed form. Substituting the explicit form of $\hat{F}_t$ (Theorem 2) into (7) and unrolling, the prior μ_t collapses into the integrand and $\hat{F}_t$ depends only on the expert predictions and observations seen so far:

$$\hat{F}_t(y) = \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_{\Theta} \exp(-2[F_t^{(p)}(y)]^2 - \frac{2}{b-a} S_{t-1}(p)) dp}{\int_{\Theta} \exp(-2[1 - F_t^{(p)}(y)]^2 - \frac{2}{b-a} S_{t-1}(p)) dp} \right), \quad (8)$$

where $S_{t-1}(p) = \sum_{j=1}^{t-1} \text{CRPS}(F_j^{(p)}, y_j)$ and dp is Lebesgue measure on Θ .

Monte Carlo implementation. The integrals over Θ are intractable in closed form but admit a simple Monte Carlo estimator. We sample $p \sim \text{Unif}(\Theta)$ via the standard exponential representation: draw $X_{nj} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ for $n \in [N], j \in [M]$ and set $v_{nj} = X_{nj} / \sum_{n'} X_{n'j}$. We reuse the same samples $\{v_{\cdot j}\}$ in both numerator and denominator: the integrands differ only in the leading squared term, so the shared $S_{t-1}(p)$ factor and correlated draws yield substantial variance reduction. This gives Algorithm 1. In practice, we assume that it is executed for M so large that the numerical integrations performed are done so with negligible error with high probability.

Algorithm 1 VT-MOE: variance-reduced Vyugin–Trunov for mixtures of experts

Require: Number of experts N , support bounds $a < b$, Monte Carlo size M .

- 1: Sample $X_{nj} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ and set $v_{nj} \leftarrow X_{nj} / \sum_{n'} X_{n'j}$ for $n \in [N], j \in [M]$.
- 2: Initialise $S_j \leftarrow 0$ for $j \in [M]$. ▷ Cumulative loss of mixture $v_{\cdot j}$.
- 3: **for** $t = 1, \dots, T$ **do**
- 4: Receive expert predictions $\{F_{n,t}\}_{n=1}^N$.
- 5: **Output** $\hat{F}_t(\cdot)$ defined for $y \in [a, b]$ by

$$\hat{F}_t(y) = \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\sum_{j=1}^M \exp(-2[(\sum_n v_{nj} F_{n,t}(y))^2 + \frac{1}{b-a} S_j])}{\sum_{j=1}^M \exp(-2[(1 - \sum_n v_{nj} F_{n,t}(y))^2 + \frac{1}{b-a} S_j])} \right).$$

- 6: Receive observation y_t .
 - 7: $S_j \leftarrow S_j + \text{CRPS}(\sum_n v_{nj} F_{n,t}, y_t)$ for each $j \in [M]$.
 - 8: **end for**
-

Theoretical Analysis

We now establish the two main theorems behind VT-MOE: that CRPS is mixable in the MOE sense (Theorem 2, justifying the closed form (8)), and that the resulting algorithm enjoys logarithmic regret with an explicit leading constant (Theorem 3). Throughout, we measure performance by the *mixture-of-experts regret*

$$\text{Reg}_T = \sum_{t=1}^T \text{CRPS}(\hat{F}_t, y_t) - \min_{p \in \Theta} \sum_{t=1}^T \text{CRPS}(F_t^{(p)}, y_t), \quad (9)$$

which, since Θ contains the vertices e_n , is at least as large as the standard best-expert regret.

E.4 MOE-mixability of CRPS

Theorem 2 (Mixability). *The CRPS loss is $\frac{2}{b-a}$ -mixable in the MOE setting. Moreover, inequality (6) is satisfied by the explicit CDF*

$$\hat{F}(y) = \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_{\Theta} \exp(-2[F^{(p)}(y)]^2) d\nu(p)}{\int_{\Theta} \exp(-2[1 - F^{(p)}(y)]^2) d\nu(p)} \right). \quad (10)$$

Proof sketch. The argument follows the structure of [24, Theorem 2] with adaptations at two key points. First, discretise $[a, b]$ at scale $\Delta = (b - a)/d$ and approximate $\text{CRPS}(F, y) \approx \Delta \sum_{s=1}^d (F(z_s) - \mathbf{1}\{z_s \geq y\})^2$ up to $O(\Delta)$ error. This step is standard. Second, where the single-expert proof applies mixability bin-by-bin to a finite expert set, we apply it bin-by-bin against the measure ν on Θ , producing $\hat{f}_s = \hat{F}(z_s)$, and then recombine bins via Hölder’s inequality with exponents aligned at $\eta = 2/(b - a)$. As noted in [9, p. 7], this amounts to replacing sums over experts by integrals over Θ with appropriate adjustments to the Hölder space. Letting $d \rightarrow \infty$ completes the proof; full details are in Appendix E.6. The rate $\eta = 2/(b - a)$ shrinks as the support widens, consistent with the intuition that prediction is harder on a larger range.

E.5 Regret bound

Theorem 3 (Regret bound). *For VT-MOE run with exact integration over Θ ($M \rightarrow \infty$),*

$$\text{Reg}_T \leq \frac{(b - a)(N - 1)}{2} \ln T + C,$$

where C depends only on a, b, N and is independent of T .

The bound is logarithmic in T and grows linearly in $N - 1$ and the support length, both of which match the intuition that larger comparator classes and wider observation ranges make prediction harder. The proof rests on two structural facts about CRPS.

Lemma 4 (Quadratic structure). *Let $L(p) = \sum_{t=1}^T \text{CRPS}(F_t^{(p)}, y_t)$, and define $A_i = \sum_t \text{CRPS}(F_{i,t}, y_t)$ and $B_{ij} = \sum_t \int_a^b (F_{i,t}(z) - F_{j,t}(z))^2 dz$. Then $L(p) = A^\top p - \frac{1}{2} p^\top B p$ for all $p \in \Theta$.*

Lemma 5 (Worst-case bounds). *For all $1 \leq i, j \leq N$ and $p \in \Theta$: (i) $|A_i - A_j| \leq 2(b - a)T$; (ii) $|(Bp)_i - (Bp)_j| \leq (b - a)T$; (iii) $-\delta^\top B \delta \leq 2(b - a)NT \|\delta\|_2^2$ for every δ with $\mathbf{1}^\top \delta = 0$.*

These bounds are worst-case and can be tightened under structural assumptions on the experts; see Section E.6.

Proof sketch. Iterating (6) and (7) yields the standard loss identity

$$\sum_{t=1}^T \text{CRPS}(\hat{F}_t, y_t) \leq \eta^{-1} \ln \frac{\nu_0(\Theta)}{\mu_{T+1}(\Theta)} = \eta^{-1} \ln \frac{1}{\mu_{T+1}(\Theta)}, \quad (11)$$

using $\mu_0 \equiv 1$. Let $p^* \in \arg \min_{\Theta} L$, which exists by continuity on the compact simplex. Combining (11) with Lemma 4 yields

$$\text{Reg}_T \leq -\frac{b-a}{2} \ln \int_{\Theta} \exp\left(-\frac{2}{b-a} [L(p) - L(p^*)]\right) dp. \quad (12)$$

Setting $\delta = p - p^*$ and using $L(p) - L(p^*) = (A - Bp^*)^\top \delta - \frac{1}{2} \delta^\top B \delta$, Lemma 5 bounds the linear term by $3(b - a)T\sqrt{N} \|\delta\|_2$ and the quadratic term by $(b - a)NT \|\delta\|_2^2$. Choosing $r = 1/(T(3\sqrt{N} + N))$ gives

$$\|\delta\|_2 \leq r \implies L(p) - L(p^*) \leq b - a \implies \exp\left(-\frac{2}{b-a} [L(p) - L(p^*)]\right) \geq e^{-2}.$$

The integral in (12) is therefore lower-bounded by $e^{-2} |\Theta \cap B(p^*, r)|$. Since Θ is a regular $(N - 1)$ -simplex, this volume is at least $c_{\Theta} r^{N-1}$ for a constant c_{Θ} depending only on N , yielding

$$\int_{\Theta} \exp\left(-\frac{2}{b-a} [L(p) - L(p^*)]\right) dp \geq C T^{-(N-1)}.$$

Substituting into (12) gives the claimed bound.

Background: Prediction with Expert Advice

Prediction with Expert Advice (PWEA) is a foundational framework in online learning and sequential decision making. In this setting, a learner repeatedly combines predictions from a set of experts and tries to perform nearly as well as the best expert in hindsight. Unlike classical statistical learning, the framework does not require assumptions about the underlying data distribution and is therefore well suited for non-stationary or adversarial environments.

Formally, let there be N experts. At each round $t \in \{1, \dots, T\}$, each expert i produces a prediction $\hat{y}_{t,i}$. The learner assigns a weight vector

$$w_t = (w_{t,1}, \dots, w_{t,N}),$$

where $w_{t,i} \geq 0$ and $\sum_{i=1}^N w_{t,i} = 1$, and forms a combined prediction. After observing the true outcome y_t , a loss function $\ell(\cdot, \cdot)$ is incurred.

The performance of the learner is typically evaluated using *regret*, defined as the difference between the cumulative loss of the learner and that of the best expert in hindsight:

$$R_T = \sum_{t=1}^T \ell_t(\hat{y}_t) - \min_{i \in \{1, \dots, N\}} \sum_{t=1}^T \ell_t(\hat{y}_{t,i}).$$

A learning algorithm is considered successful if its regret grows sublinearly with time, i.e.,

$$R_T = o(T),$$

which implies that the average regret vanishes asymptotically.

One of the most influential algorithms in this framework is the Exponentially Weighted Average (EWA) forecaster, also known as Hedge. The method updates expert weights according to their past losses:

$$w_{t+1,i} = \frac{w_{t,i} \exp(-\eta \ell_t(\hat{y}_{t,i}))}{\sum_{j=1}^N w_{t,j} \exp(-\eta \ell_t(\hat{y}_{t,j}))},$$

where $\eta > 0$ is the learning rate. Experts with lower historical loss receive larger weights over time.

A central concept in the theory is that of *mixable losses*. For mixable loss functions, algorithms based on exponential weighting admit strong logarithmic regret guarantees of the form

$$R_T \leq \frac{\log N}{\eta},$$

which are independent of the time horizon T . Common examples include the logarithmic loss and squared loss under suitable assumptions.

Prediction with Expert Advice has been widely applied in ensemble forecasting, finance, adaptive control, and weather prediction, where multiple models exhibit varying performance across spatial and temporal regimes. The framework is particularly attractive in operational forecasting settings because it enables adaptive combination of heterogeneous models while retaining theoretical guarantees on long-term performance.

E.6 Vovk's aggregation algorithm

Vovk's exponentially weighted forecaster [23] maintains weights $w_t \in \Theta$, initialised to the uniform vector $w_1 = (1/N, \dots, 1/N)$. At each round it outputs the linear mixture $\hat{F}_t = \sum_n (w_t)_n F_{n,t}$, then updates

$$(w_{t+1})_n = \frac{(w_t)_n e^{-\text{CRPS}(F_{n,t}, y_t)}}{\sum_m (w_t)_m e^{-\text{CRPS}(F_{m,t}, y_t)}}.$$

See Algorithm 2.

Algorithm 2 Vovk's aggregation algorithm

Require: Number of experts N , support bounds $a < b$.

- 1: Initialise $(w_1)_n \leftarrow 1/N$ for $n \in [N]$.
 - 2: **for** $t = 1, 2, \dots, T$ **do**
 - 3: Receive expert predictions $\{F_{n,t}\}_{n=1}^N$.
 - 4: **Output** $\hat{F}_t = \sum_{n=1}^N (w_t)_n F_{n,t}$.
 - 5: Receive observation y_t .
 - 6: Update $(w_{t+1})_n \leftarrow \frac{(w_t)_n e^{-\text{CRPS}(F_{n,t}, y_t)}}{\sum_m (w_t)_m e^{-\text{CRPS}(F_{m,t}, y_t)}}$ for each $n \in [N]$.
 - 7: **end for**
-

For the VT-MOE algorithm, the integral over $\Theta = \Delta^{N-1}$ is approximated using Monte Carlo integration with M simplex samples drawn once and reused across rounds (Common Random Numbers).

Proof of Mixability

In this section, we furnish the complete proof of Theorem 2. In order to do this, we require the following result which is informally stated in [9], and is a continuous version of the analogous result in [13]. The proof is a direct calculation which we will omit.

Proposition 6. *Consider the loss function $l' : \{0, 1\} \times [0, 1] \rightarrow [0, 1]$ given by $l'(\omega, y) = (\omega - y)^2$. Let π be any finite measure on $[0, 1]$. Then, there exists $\phi \in [0, 1]$ such that*

$$e^{-2l'(\phi, y)} \leq \frac{\int_0^1 e^{-2l'(t, y)} d\pi(t)}{\pi([0, 1])} \text{ for } y \in \{0, 1\}.$$

An explicit formula for ϕ is given by

$$\phi = \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_0^1 e^{-2l'(t, 0)} d\pi(t)}{\int_0^1 e^{-2l'(t, 1)} d\pi(t)} \right).$$

We are now ready to begin the proof.

Proof of Theorem 2. We shall borrow several notations and arguments from the proof of [24, Theorem 2].

Suppose that $F_n, n \in [N]$ are expert predictions and ν is a finite measure over Θ .

For any CDF F of a random variable supported on $\mathcal{X}$ and $d \in \mathbb{N}$, define $z_k = \frac{k}{d}a + \frac{d-k}{d}b$ for $k = 0, \dots, d$. Note that $z_0 = a, z_d = b$ and z_k are equally spaced in between $b - a$ having distance $\Delta = \frac{b-a}{d}$ between consecutive points. Now, define the CDF F_d by

$$F_d(z) = \begin{cases} 0 & z \leq a \\ 1 & z \geq b \\ F(z_k) & z \in [z_{k-1}, z_k], 1 \leq k \leq d. \end{cases} \quad (13)$$

One may consider F_d as a "discretization" of F . Now, we can derive the bound below exactly as in [24, equation (21)].

$$|\text{CRPS}(F_d, y) - \text{CRPS}(F, y)| \leq 2\Delta \quad (14)$$

for all $y \in \mathbb{R}$.

We will perform yet another simplification. For $y \in \mathbb{R}$ and $s \in [d]$, let $\omega_{y,s} = 1_{\{z_s \geq y\}} \in \{0, 1\}$ and $\omega_y = (\omega_{y,1}, \dots, \omega_{y,d})$. Roughly, ω_y is a discretization of the CDF $1_{\{z \leq y\}}$.

Once again, exactly as in [24, equation (22)], we obtain

$$\left| \text{CRPS}(F_d, y) - \Delta \sum_{s=1}^d (F(z_s) - \omega_{y,s})^2 \right| \leq \Delta$$

for all $y \in \mathbb{R}$. Combining this with (14) we have

$$\left| \Delta \sum_{s=1}^d (F(z_s) - \omega_{y,s})^2 - \text{CRPS}(F, y) \right| \leq 3\Delta$$

for all $y \in \mathbb{R}$. Consequently, for any CDF F supported on $[a, b]$ and $y \in \mathbb{R}$,

$$\lim_{d \rightarrow \infty} \Delta \sum_{s=1}^d (F(z_s) - \omega_{y,s})^2 = \text{CRPS}(F, y). \quad (15)$$

Having completed this approximation, we return to the expert predictions $F_n, n \in [N]$ and measure ν over Θ . For ease, we let $p = (p_1, \dots, p_N) \in \Theta$ denote a probability vector in the simplex.

Recall that our aim is to find a CDF $\hat{F}$ such that

$$e^{-\frac{2}{b-a} \text{CRPS}(\hat{F}, y)} \geq \frac{\int_{\Theta} e^{-\frac{2}{b-a} \text{CRPS}(\sum_{n=1}^N p_n F_n, y)} d\nu(p)}{\nu(\Theta)} \text{ for all } y \in \mathbb{R}.$$

Fix $1 \leq s \leq d$. Define the measure π_s on $[0, 1]$ by

$$\pi_s(A) = \frac{\int_{\Theta} 1_{\{\sum_{n=1}^N p_n F_n(z_s) \in A\}} d\nu(p)}{\nu(\Theta)}. \quad (16)$$

Observe that $\pi_s([0, 1]) = 1$. Now, there exists $\hat{f}_s \in [0, 1]$ such that

$$e^{-2l'(\hat{f}_s, \Omega_s)} \geq \int_0^1 e^{-2l'(t, \Omega_s)} d\pi_s(t) \quad (17)$$

for all $\Omega_s \in \{0, 1\}$, where $\hat{f}_s = \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_0^1 e^{-2l'(t, 0)} d\pi_s(t)}{\int_0^1 e^{-2l'(t, 1)} d\pi_s(t)} \right)$ by Proposition 6. By the definition (16) of π_s and Fubini's theorem we have

$$\begin{aligned} \int_0^1 e^{-2l'(t, \Omega_s)} d\pi_s(t) &= \frac{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \\ \hat{f}_s &= \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), 0)} d\nu(p)}{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), 1)} d\nu(p)} \right). \end{aligned} \quad (18)$$

Combining the first of these with (17),

$$e^{-2l'(\hat{f}_s, \Omega_s)} \geq \frac{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)}.$$

Multiplying this estimate over $s \in [d]$ and taking both sides to the power Δ gives

$$e^{-2\Delta \sum_{s=1}^d l'(\hat{f}_s, \Omega_s)} \geq \prod_{s=1}^d \left(\frac{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \right)^{\Delta}. \quad (19)$$

Let $h_s(p) = e^{-2\Delta l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)}$ for $p \in \Theta$. Observe that by Hölder's inequality,

$$\prod_{s=1}^d \|h_s\|_{\frac{1}{\Delta}} \geq \left\| \prod_{s=1}^d h_s \right\|_{\frac{1}{\Delta d}}.$$

Here, all norms are with respect to the measure ν on Θ . Therefore, recalling the definition of h_s we have

$$\prod_{s=1}^d \left(\frac{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \right)^{\Delta} \geq \left(\frac{\int_{\Theta} e^{-\frac{2}{d} \sum_{s=1}^d l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \right)^{\Delta d}.$$

Combining this with (19) and noting that $\Delta d = b - a$ we have

$$e^{-2\Delta \sum_{s=1}^d l'(\hat{f}_s, \Omega_s)} \geq \left(\frac{\int_{\Theta} e^{-\frac{2}{b-a} \Delta \sum_{s=1}^d l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \right)^{b-a}.$$

A simple rearrangement yields the bound

$$e^{-\frac{2}{b-a} \Delta \sum_{s=1}^d l'(\hat{f}_s, \Omega_s)} \geq \frac{\int_{\Theta} e^{-\frac{2}{b-a} \Delta \sum_{s=1}^d l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)}. \quad (20)$$

Finally, let $y \in \mathbb{R}$ be arbitrary. Apply the above estimate to $\Omega_s = (\omega_{y,s})$ to obtain

$$e^{-\frac{2}{b-a} \Delta \sum_{s=1}^d (\hat{f}_s - \omega_{y,s})^2} \geq \frac{\int_{\Theta} e^{-\frac{2}{b-a} \Delta \sum_{s=1}^d l'(\sum_{n=1}^N p_n F_n(z_s), \omega_{y,s})} d\nu(p)}{\nu(\Theta)}, \quad (21)$$

where by (18) and the definition of l' ,

$$\begin{aligned} \hat{f}_s &= \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), 0)} d\nu(p)}{\int_{\Theta} e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), 1)} d\nu(p)} \right) \\ &= \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_{\Theta} e^{-2[\sum_{n=1}^N p_n F_n(z_s)]^2} d\nu(p)}{\int_{\Theta} e^{-2[1 - \sum_{n=1}^N p_n F_n(z_s)]^2} d\nu(p)} \right). \end{aligned}$$

In particular, we have that $\hat{f}_s = \hat{F}(z_s)$ where

$$\hat{F}(y) = \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_{\Theta} e^{-2[\sum_{n=1}^N p_n F_n(y)]^2} d\nu(p)}{\int_{\Theta} e^{-2[1 - \sum_{n=1}^N p_n F_n(y)]^2} d\nu(p)} \right)$$

is a well-defined CDF. Finally, letting $d \rightarrow \infty$ in (21), by Fubini's theorem and (15) we obtain

$$e^{-\frac{2}{b-a} \text{CRPS}(\hat{F}, y)} \geq \frac{\int_{\Theta} e^{-\frac{2}{b-a} \text{CRPS}(\sum_{n=1}^N p_n F_n, y)} d\nu(p)}{\nu(\Theta)}.$$

This completes the proof. $\square$

Proof of Regret Bound

In this section, we prove the regret bound Theorem 3. First, we shall assume the truth of Lemmas 4 and 5, and prove the theorem. Subsequently, we shall prove these additional lemmas.

Proof of Theorem 3. Let $p^* \in \arg \min_{\Theta} \sum_{t=1}^T \text{CRPS} \left(\sum_{n=1}^N p_n F_{n,t}, y_t \right)$. This exists, since we are minimizing a continuous function over a compact set.

Iterating the equation (7) over $t \in [T]$ and applying Lemma 4 we obtain

$$\mu_{T+1}(\Theta) = \int_{\Theta} e^{-\frac{2}{b-a} \sum_{t=1}^T \text{CRPS}(\sum_{n=1}^N p_n F_{n,t}, y_t)} dp = \int_{\Theta} e^{-\frac{2}{b-a} (A^T p - p^T B p)} dp,$$

where the integration is now performed with respect to the $(N-1)$ -dimensional Lebesgue measure on Θ . Recall that $p \in \Theta$ is written as $p = (p_1, \dots, p_N)$.

Now, by (11) and Lemma 4,

$$\begin{aligned} \text{Reg}_{M,T} &\leq -\frac{b-a}{2} \ln \mu_{T+1}(\Theta) - \sum_{t=1}^T \text{CRPS} \left(\sum_{n=1}^N p_n^* F_{n,t}, y_t \right) \\ &= -\frac{b-a}{2} \ln \int_{\Theta} e^{-\frac{2}{b-a} (A^T p - \frac{1}{2} p^T B p)} - \left[A^T p^* - \frac{1}{2} (p^*)^T B p^* \right] \\ &= -\frac{b-a}{2} \ln \int_{\Theta} e^{-\frac{2}{b-a} (A^T p - \frac{1}{2} p^T B p)} dp - \frac{b-a}{2} \ln e^{\frac{2}{b-a} (A^T p^* - \frac{1}{2} (p^*)^T B p^*)} \\ &= -\frac{b-a}{2} \ln \int_{\Theta} e^{\frac{2}{b-a} [A^T p^* - \frac{1}{2} (p^*)^T B p^* - (A^T p - \frac{1}{2} p^T B p)]} dp. \end{aligned}$$

Let $L(p) = \sum_{t=1}^T \text{CRPS}(\sum_{n=1}^N p_n F_{n,t}, y_t)$. Then, we have shown that

$$\text{Reg}_{M,T} \leq -\frac{b-a}{2} \ln \int_{\Theta} e^{-\frac{2}{b-a} (L(p_1, \dots, p_N) - L(p^*))} dp. \quad (22)$$

We shall now further simplify the exponent. To observe this, we shall simply treat L as being expanded in a Taylor series about p^* . Note that $\nabla L(p^*) = A - B p^*$, while the Hessian $\nabla^2 L(p^*) = -B$. Since L is quadratic, there are no further terms to consider. Therefore, if $\delta = (p - p^*)$, then the Taylor expansion yields

$$L(p) - L(p^*) = (A - B p^*)^T \delta - \frac{1}{2} \delta^T B \delta. \quad (23)$$

Observe that we do not assume anything about the point p^* at all during this manipulation: indeed, if p^* is interior to Θ then $(A - B p^*) = 0$ since p^* is an interior local minima of L in Θ . However, if p^* is on the boundary, then we cannot assume that the derivative is zero at p , since L may decrease in directions away from Θ at p . Note that $\mathbf{1}^T \delta = 0$ where $\mathbf{1}$ is a vector all of whose entries are 1.

We can now attempt to bound the right hand side of (23). By Lemma 5,

$$\begin{aligned} (A - B p^*)^T \delta &= (A - B p^* + \min_i (A - B p^*)_i \mathbf{1})^T \delta \\ &\leq \max_{1 \leq i, j \leq N} |(A_i - (B p^*)_i) - (A_j - (B p^*)_j)| \times \|\delta\|_1 \\ &\leq \left[\max_{1 \leq i, j \leq N} |A_i - A_j| + \max_{1 \leq i, j \leq N} |(B p^*)_i - (B p^*)_j| \right] \times \|\delta\|_1 \\ &\leq [2(b-a)T + (b-a)T] \times \|\delta\|_1 \\ &\leq 3(b-a)T \|\delta\|_1 \leq 3(b-a)T \sqrt{N} \|\delta\|_2, \end{aligned}$$

where we used the Cauchy-Schwarz inequality in the last line. For the other term, we simply have

$$-\frac{1}{2} \delta^T B \delta \leq (b-a)NT \|\delta\|_2^2$$

by Lemma 5. Combining the above two bounds and (23),

$$L(p) - L(p^*) \leq 3(b-a)T \sqrt{N} \|\delta\|_2 + (b-a)NT \|\delta\|_2^2 \leq (b-a)T(3\sqrt{N} + N) \|\delta\|_2 \quad (24)$$

since $\|\delta\|_2 \leq 1$. Now, let $r = \frac{1}{T(3\sqrt{N} + N)}$. Then, if $\|\delta\|_2 \leq r$, applying the definition of L and (24) implies that $L(p) - L(p^*) \leq b-a$. In particular,

$$\|p^* - p\| \leq r \implies e^{-\frac{2}{b-a} (L(p) - L(p^*))} \geq e^{-2}.$$

Finally, let $S = \Theta \cap \{\|p^* - p\| \leq r\}$. Then,

$$\int_{\Theta} e^{-\frac{2}{b-a}(L(p_1, \dots, p_N) - L(p^*))} dp \geq e^{-2}|S|,$$

where $|S|$ is the $(N-1)$ -dimensional Lebesgue measure of S . This combined with (22) yields

$$\text{Reg}_{M,T} \leq -\frac{b-a}{2} \ln e^{-2}|S| \leq \frac{b-a}{2} \ln \frac{1}{|S|} + (b-a). \quad (25)$$

It remains to bound $|S|$. However, a property of the simplex Θ is that for some universal constant c_{Θ} depending upon only n , $|\Theta \cap \mathbf{B}(q, s)| \geq c_{\Theta} |\mathbf{B}(q, s)|$ for all $q \in \Theta, s > 0$, where $\mathbf{B}(q, s) = \{q' : \|q - q'\| \leq s\}$. Thus,

$$\ln \frac{1}{|S|} \leq \ln \frac{1}{c_{\Theta} |\mathbf{B}(p^*, r)|} = (N-1) \ln \frac{1}{r} + C = (N-1) \ln(T) + C$$

where C is independent of T . Finally, combining this with (25) gives

$$\text{Reg}_{M,T} \leq \frac{(b-a)(N-1)}{2} \ln(T) + C$$

where C has no dependence on T , completing the proof. $\square$

Now, we shall prove the lemmas.

Proof of Lemma 4. Let $t \in [T]$, $p \in \Theta$ and $x, y \in [a, b]$ be arbitrary. Let $F_{n,t}, n \in [N]$ denote the expert predictions at time t . We have

$$\begin{aligned} (F_t^{(p)}(x) - 1_{\{x \leq y\}})^2 &= \left(\sum_{n=1}^N p_n F_{n,t}(x) - 1_{\{x \leq y\}} \right)^2 \\ &= \left(\sum_{n=1}^N p_n F_{n,t}(x) \right)^2 - 2 \cdot 1_{\{x \leq y\}} \sum_{n=1}^N p_n F_{n,t}(x) + 1_{\{x \leq y\}}. \end{aligned}$$

On the other hand,

$$\sum_{n=1}^N p_n (F_{n,t}(x) - 1_{\{x \leq y\}})^2 = \sum_{n=1}^N p_n F_{n,t}(x)^2 + 1_{\{x \leq y\}} - 2 \sum_{n=1}^N p_n F_{n,t}(x) 1_{\{x \leq y\}}.$$

Subtracting the previous equation from this one,

$$\sum_{n=1}^N p_n (F_{n,t}(x) - 1_{\{x \leq y\}})^2 - (F_t^{(p)}(x) - 1_{\{x \leq y\}})^2 = \sum_{n=1}^N p_n F_{n,t}(x)^2 - \left(\sum_{n=1}^N p_n F_{n,t}(x) \right)^2.$$

However, we note the algebraic identity

$$\sum_{n=1}^N p_n F_{n,t}(x)^2 - \left(\sum_{n=1}^N p_n F_{n,t}(x) \right)^2 = \frac{1}{2} \sum_{n,m=1}^N p_m p_n (F_{n,t}(x) - F_{m,t}(x))^2.$$

Combining this with the previous equality,

$$(F_t^{(p)}(x) - 1_{\{x \leq y\}})^2 = \sum_{n=1}^N p_n (F_{n,t}(x) - 1_{\{x \leq y\}})^2 - \frac{1}{2} \sum_{n,m=1}^N p_m p_n (F_{n,t}(x) - F_{m,t}(x))^2.$$

Summing this identity over $t \in [T]$ and then integrating over $x \in \mathbb{R}$, the identity immediately follows once we recall the definitions of A and B from the statement of the lemma. $\square$

Next, we prove Lemma 5.

Proof of Lemma 5. Note that if F, G are CDFs of random variables concentrated on $[a, b]$, then

$$\int_a^b (F(x) - G(x))^2 dx \leq \int_a^b 1 dx \leq (b - a). \quad (26)$$

From here, part (1) follows directly from the triangle inequality $|A_i - A_j| \leq |A_i| + |A_j|$ and the definition of A .

For the proof of part (2), let $p = (p_1, \dots, p_N) \in \Theta$ be arbitrary. Then, for any $i, j \in [N]$,

$$|(Bp)_i - (Bp)_j| \leq \sum_{n=1}^N |B_{ni}p_i - B_{nj}p_j| \leq \max_{n,k \in [N]} |B_{nk}|$$

since $p_i, p_j \in [0, 1]$ and B has non-negative entries. However, this quantity is bounded by $T(b - a)$ by (26). The result follows.

Finally, for part (3) let $\delta = (\delta_1, \dots, \delta_N)$ be an arbitrary vector orthogonal to $\mathbf{1}$ (so that its entries sum to 0). One can verify that

$$\begin{aligned} -\delta^T B \delta &= -\sum_{t=1}^T \sum_{i,j=1}^N \delta_i \delta_j \int_a^b (F_{i,t}(z) - F_{j,t}(z))^2 dz \\ &= 2 \sum_{t=1}^T \int_a^b \left(\sum_{n=1}^N \delta_n F_{n,t}(z) \right)^2 dz. \end{aligned}$$

Following this, we note that $\left(\sum_{n=1}^N \delta_n F_{n,t}(z) \right)^2 \leq N \|\delta\|_2^2$ by the Cauchy-Schwarz inequality. Applying this to the previous inequality, we obtain

$$-\delta^T B \delta \leq 2 \sum_{t=1}^T \int_a^b \|\delta\|_2^2 dz = 2(b - a)NT\|\delta\|_2^2.$$

Thus, the proofs of all three parts are complete. $\square$

We end this section with a detailed comparison with [9, Theorem 4]. As mentioned in Section E.3.1, the optimal rate obtained by them is $\frac{1}{4} \ln T$, while we obtain an extra factor of 2 leading to the apparently non-optimal $\frac{1}{2} \ln T$.

This non-optimality arises from the term $(A - Bp^*)$ in (24). Indeed, were this term either zero or smaller in order, then $L(p) - L(p^*)$ could be bounded by $CT\|\delta\|_2$ where C is independent of T , resulting in an improvement by a factor of $\frac{1}{2}$.

The naive expert assumption affords us this stronger bound as follows. Recall that in the naive expert setting, $a = 0, b = 1, N = 2$ and $F_{0,t} = 1_{\{x \leq 0\}}$ while $F_{1,t} = 1_{\{x \leq 1\}}$. Let $y_i, i \in [T]$ be the observations up till time T . One easily computes from (4) that

$$A = \left(\sum_{t=1}^T y_t, T - \sum_{t=1}^T y_t \right)$$

and

$$B = \begin{pmatrix} 0 & T \\ T & 0 \end{pmatrix}.$$

We now compute p^* . Since we are in the setting of two experts, any probability vector is of the form $(1 - q, q)$ for $q \in [0, 1]$. Now, by Lemma 4 and the above formulas we have

$$L(1 - q, q) = (1 - 2q) \sum_{t=1}^T y_t + Tq^2.$$

Minimizing this over q by taking the derivative yields $q^* = \frac{\sum_{t=1}^T y_t}{T}$, which is nothing but the empirical average of the observations y_t . Thus, $p^* = (1 - q^*, q^*)$.

At this point, some simple algebra yields $A - Bp^* = 0$. Following the proof above as usual does yield the optimal bound $\frac{1}{4}$ for [9], demonstrating that (up to the additive constant) we achieve minimax-optimal regret for naive two-expert predictions on $[0, 1]$.

Another way of explaining this is as follows. Recall that in the proof of Theorem 3, we were unable to set $(A - Bp^*) = 0$, since p^* could lie on the boundary. However if B is invertible and $B^{-1}A$ is a well-defined probability vector, then $p^* = B^{-1}A$ satisfies $(A - Bp^*) = 0$ and is a typical critical point of L , reducing the bound by a factor of 2. This is precisely what happens in [9] and would also be expected to occur in a number of situations, including when p^* is guaranteed to strictly lie inside the simplex.