Learning Observation-Guided Probabilistic Forecast Mixtures with Online Regret Guarantees

Anonymous authors
Paper under double-blind review

Abstract

Data-driven weather models have achieved similar overall performance to physics-based methods, but no single model wins benchmarks. A common next step is blending them, but existing methods are usually simple averages over coarse boxes with less online adaptation and no theoretical relative error guarantees. We look at historical forecast ground *station* pairs and learn the best *probabilistic* mixed model through supervised learning. We then adapt online to streamed observations with expert advice algorithms, and develop novel relative error bounds in terms of probabilistic ensemble metrics. We get 30% + improvements in evaluations relative to standard methods and lower error relative to estimates of the best possible forecast in hindsight, in evaluations over multiple variables and regions. This is extensible to any observation set which can help create bespoke forecasts customized to different uses with formal guarantees.

1 Introduction & Context

Weather forecasting is essential for many fields like agriculture, energy and disaster management. Neural weather models like FourCastNet (Pathak et al., 2022), GenCast (Price et al., 2024), and FGN (Alet et al., 2025) have become compute-efficient competitors to traditional numerical weather predictors (NWPs) on various metrics Rasp et al. (2024). With diverse architectures and biases, these models show similar aggregate performances, with no one model dominating benchmarks on a consistent basis across region and variables. Some natural approaches to combine them include offline methods, which learn a combiner based on historical forecast-truth data. These are data driven by using 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 (Zhang et al., 2023; Lau et al., 2025; Fu et al., 2022; Liu & Etesami, 2025). NWP statistical post-processing methods like ensemble model output statistics (EMOS) (Gneiting et al., 2005), Bayesian model averaging (BMA) (Raftery et al., 2005), and proper-scoring-rule calibration (Gneiting & Raftery, 2007) 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 Vovk (1990; 1995); Freund (1996); Littlestone & Warmuth (1994); Cesa-Bianchi & Lugosi (2006). 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 Chang & Sapsis (2026).

Our extensible framework learns the best *probabilistic* forecast from historical observations, then cheaply adapts online to data as it comes along. Moreover, using the machinery of expert advice methods, we provide a novel formal bound that gives a guarantee on the relative error or the *regret* in terms of the Continuous Ranked Probability Score (CRPS) of the final probabilistic forecast relative to the best forecast or possible forecast combination in hindsight, along with a first evaluation of neural models on the Automated Weather Stations (AWS) network of the Indian Meteorological Department (IMD). Around **40 %**

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.

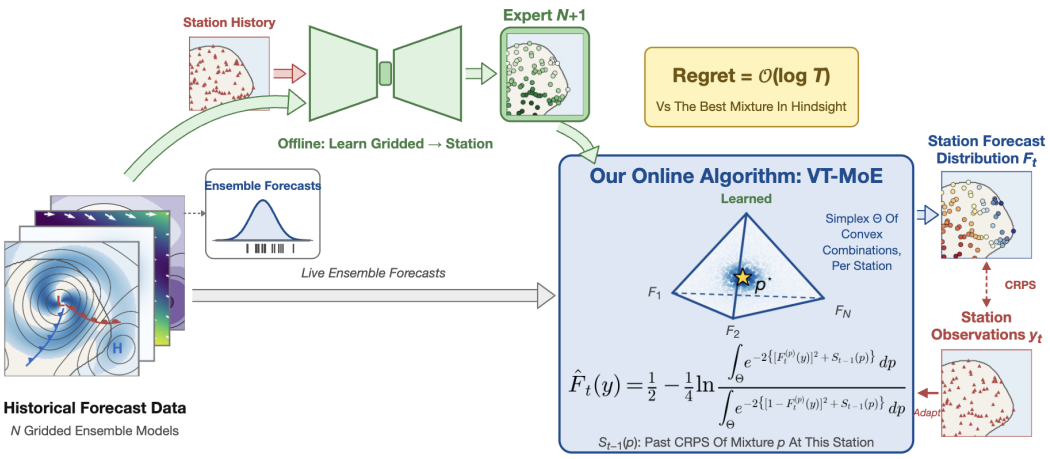


Figure 1: Outline of our Framework

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).

This work can be extended to real world impact since 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.

We now outline the structure of this article. In Section 2 we formulate the problem and state the main theoretical result, Theorem 1. In Section 3 we detail the data and experiments which substantiate our theoretical results. We discuss the results in Section 4, and conclude in Section 5.

2 Problem Formulation & Theoretical Results

Forecasts and observations. We work in the probabilistic forecasting setup with multiple forecast trajectories generated for each day by each model. 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 station field $y_{t,\tau} \in \mathcal{Y} \subseteq \mathbb{R}^{|\mathcal{G}|}$ on a spatial grid $\mathcal{G}$, with each grid-point value lying in a known interval $[a, b] \subset \mathbb{R}$. 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.

Loss. Quality is measured by the *continuous ranked probability score* (CRPS) (Gneiting & Raftery, 2007), evaluated pointwise and averaged over $\mathcal{G}$. For any predictive CDF $\hat{F}$ and observation y at a single grid point,

$$\mathrm{CRPS}(\hat{F}, y) \;=\; \int_a^b \big(\hat{F}(z) - \mathbf{1}\{z \geq y\}\big)^2 \, dz. \tag{1}$$

For an empirical CDF given by an ensemble $\{e_1, \ldots, e_M\}$, CRPS admits the closed form $\frac{1}{M}\sum_i |e_i - y| - \frac{1}{2M^2}\sum_{i,j} |e_i - e_j|$. In implementation we use the unbiased $M(M-1)$-normalised plug-in estimator (details in Appendix). We let CRPS act on grid-valued

$(\hat{F}_{t,\tau}, y_{t,\tau})$ by averaging (1) over $g \in \mathcal{G}$, and write the cumulative loss as

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

CRPS is a strictly proper scoring rule (Gneiting & Raftery, 2007) and is $\frac{2}{b-a}$-mixable on $[a, b]$ (V'yugin & Trunov, 2021; Dombry & Zaoui, 2024). It also carries a quadratic structure (Lemma 4) that is central to controlling its fluctuations. Both properties drive the logarithmic regret bound.

We now define our notion of regret. Let

$$\Theta = \left\{ (w_1, \ldots, w_N) : w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\} \tag{2}$$

be the simplex of convex mixtures of N experts. For a fixed time $t \in [T]$, and $p = (p_1, \ldots, p_n) \in \Theta$ define the mixture CDF $F^p(t) = \sum_{i=1}^{n} p_i F_{i,t}$. Informally, the mixture CDF randomly chooses to predict using the ith expert's CDF, with probability p_i.

Define the accumulated loss until time T by

$$L_{w,T} = \sum_{t \leq T} \mathrm{CRPS}(F_t^{(p)}, y_t), \quad L_T = \sum_{t \leq T} \mathrm{CRPS}(\hat{F}_t, y_t). \tag{3}$$

Now, we define the *regret over best mixture in hindsight* by

$$\mathrm{Reg}_T = L_T - \min_{p \in \Theta} L_{w,T}. \tag{4}$$

Note that the regret depends on the algorithm we use to combine the expert predictions.

Our algorithm, in this article, is fairly simple in that our prediction $\hat{F}_t$ can explicitly be calculated based on the expert predictions and observations seen so far:

$$\hat{F}_t(y) \;=\; \tfrac{1}{2} \;-\; \tfrac{1}{4} \ln\left(\frac{\int_\Theta \exp\big(-2[F_t^{(p)}(y)]^2 + \frac{2}{b-a} S_{t-1}(p) \big)\, dp}{\int_\Theta \exp\big(-2[1 - F_t^{(p)}(y)]^2 + \frac{2}{b-a} S_{t-1}(p) \big)\, dp} \right), \tag{5}$$

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

Algorithm 1 Theoretical Mixture-of-Experts Algorithm

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

1: Initialize $S_0(p) \leftarrow 0$ for $p \in \Theta$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Receive expert predictions $\{F_{n,t}\}_{n=1}^{N}$.
4: **Output** Predict $\hat{F}_t(\cdot)$ defined by

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

5: Receive observation y_t.
6: $S_t(p) \leftarrow S_{t-1}(p) + \mathrm{CRPS}(F_t^{(p)}, y_t)$ for each $p \in \Theta$.
7: **end for**

Our main theoretical result is a regret bound for a version of this algorithm that takes into account the need to numerically integrate the two functions over the set Θ, due to the absence of an explicit anti-derivative.

We use a standard Monte-Carlo approximation to do this. Namely, fix a large integer $\mathsf{M} \in \mathbb{N}$. We sample $p \sim \mathrm{Unif}(\Theta)$ via the standard exponential representation: draw $X_{nj} \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$

for $n \in [N], j \in [\mathsf{M}]$ 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 2. 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 2 VT-MOE: variance-reduced Vyugin–Trunov for mixtures of experts

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

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

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

  Receive observation y_t.
  $S_j \leftarrow S_j + \text{CRPS}\big(\sum_n v_{nj} F_{n,t},\, y_t \big)$ for each $j \in [\mathsf{M}]$.
end for

Now we state the regret bound for Algorithm 2 which reduces to one for Algorithm 1.

Theorem 1 (Regret bound)**.** *Fix $T \in \mathbb{N}$. For any sequence of observations $y_t, t \in [T]$ and expert predictions $F_{i,t}, i \in [N], t \in [T]$, the random regret $\text{Reg}_{\mathsf{M},T}$ incurred by implementing Algorithm 2 with parameter $\mathsf{M} > 0$ satisfies*

$$\mathbb{P}\left(\sqrt{\text{Reg}_{\mathsf{M},T}} \;\le\; \sqrt{\frac{(b-a)(N-1)}{2} \ln T \;+\; C_1} + \epsilon \right) \ge 1 - \frac{4C_2(b-a)}{\mathsf{M} \min\{\epsilon^2, \frac{\epsilon^4}{4T(b-a)}\}} T^2 e^{4T}. \quad (6)$$

for any $\epsilon > 0$, where C_1 depends only on a, b, N (in particular, it is independent of T), and C_2 is an independent constant. Similarly, for Algorithm 2 we have the regret bound:

$$\text{Reg}_T \le \frac{(b-a)(N-1)}{2} \ln T + C. \quad (7)$$

Furthermore, if any minimizer $p^* \in \arg\min_{p \in \Theta} L_{p,T}$ lies in the interior of Θ (i.e. $p_i^* > 0$ for $i = 1, \ldots, N$), then we obtain the better regret bounds

$$\mathbb{P}\left(\sqrt{\text{Reg}_{\mathsf{M},T}} \;\le\; \sqrt{\frac{(b-a)(N-1)}{4} \ln T \;+\; C_1} + \epsilon \right) \ge 1 - \frac{4C_2(b-a)}{\mathsf{M} \min\{\epsilon^2, \frac{\epsilon^4}{4T(b-a)}\}} T^2 e^{4T}, \text{ and} \quad (8)$$

$$\text{Reg}_T \le \frac{(b-a)(N-1)}{4} \ln T + C_1. \quad (9)$$

where the value of C_1, C_2 are as in (6)*.*

The leading term $\frac{(b-a)(N-1)}{4} \ln T$ is minimax optimal. Indeed, for the particular case $a = 0, b = 1$, with $y_t \in \{0, 1\}$ for $t \in [T]$ and $N = 2$ experts making the constant predictions $F_{1,t}(x) = \mathbf{1}_{x \ge 0}$ and $F_{2,t}(x) = \mathbf{1}_{x \ge 1}$, it was proved in (Freund, 1996) that Algorithm 1 attains a tight regret bound of $\frac{1}{4} \ln T$, which falls under the special case of the theorem above. We provide more details on this in the appendix.

The proof of Theorem 1 can be found in Appendix E. We will now describe the datasets and experiments used for our empirical results.

3 Data and Experiments

Forecast Data: 2019-2025 range (post ML models' training period) is taken from FourCastNet-3, GenCast AI weather models, 2022-2025 from FGN, and IFS-ENS, GEM,

GEFS Numerical Weather Predictors, 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 $\tau \in \mathcal{T} = \{12, \ldots, 72\}$ 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: Our offline aggregator is a U-Net that maps expert forecasts to a probabilistic mixture forecast over the target ground truth at L lead times. It learns which forecast models to weigh more based on historical data. More details are in the Appendix.

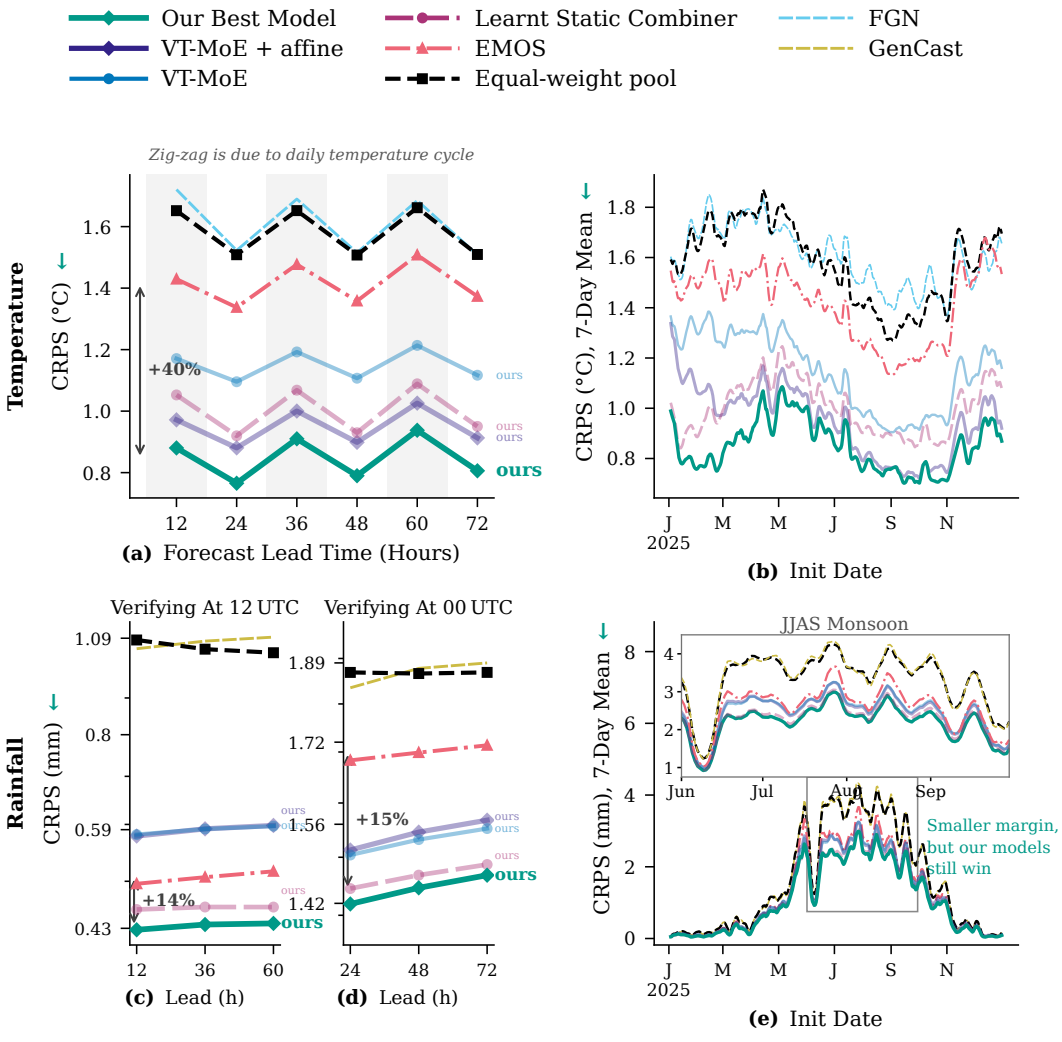


Figure 2: **Temperature, rainfall India CRPS:** (a, c) CRPS avg per forecast lead time (2024-2025); (b, e) the daily 7-day mean sample rollout just for 2025. The zig-zag in (a) due to the leads 12/36/60 h verifying at 12 UTC (shaded) and 24/48/72 h at 00 UTC

Online Adaptation 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). Alongside the previous theoretically guaranteed VTMOE approach in Algorithm

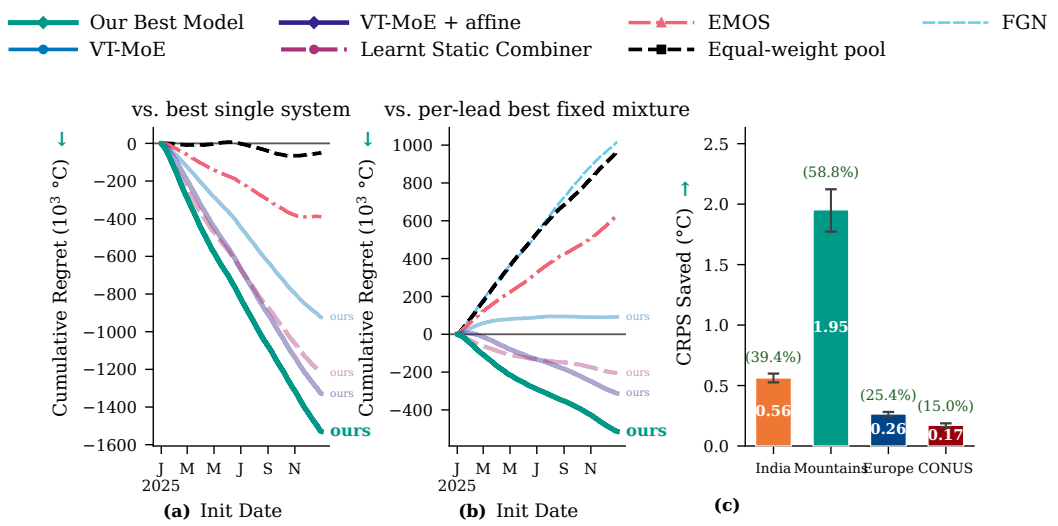


Figure 3: **Regret:** (a)(b) for temperature (rainfall in appendix), learner adds affine correction so lies outside comparator class hence giving negative regret; (c) shows regionwise variation in saved temperature CRPS relative to EMOS

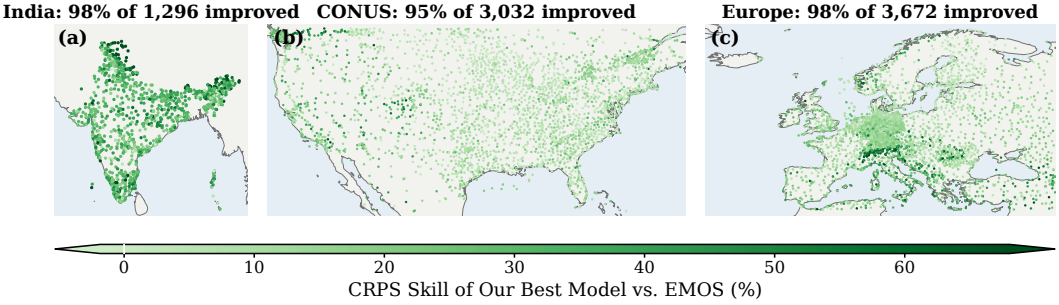


Figure 4: **Improvement over standard EMOS method**

2, we also have an affine per-station correction to it, and a learnt online combiner. In this combiner, for a set A_r of experts, it keeps three parameters per key(a key being a (station, lead) pair), convex weights $w_i \geq 0, i = 1, \ldots, |A_r|$ satisfying $\sum_{i=1}^{|A_r|} w_i = 1$, an additive shift b and a log-scale u. It divides the experts' ensembles as quantiles and pools them level by level, followed by recalibration 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. 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. It is to note that while the theoretically guaranteed algorithm doesn't have the best empirical performance, it still improves upon a standard bias correcting method and the equal weighted pool.

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

4 Results

Figure 2 shows our model having lower CRPS than other methods for both temperature and precipitation, with the improvement being relatively larger for temperature due to

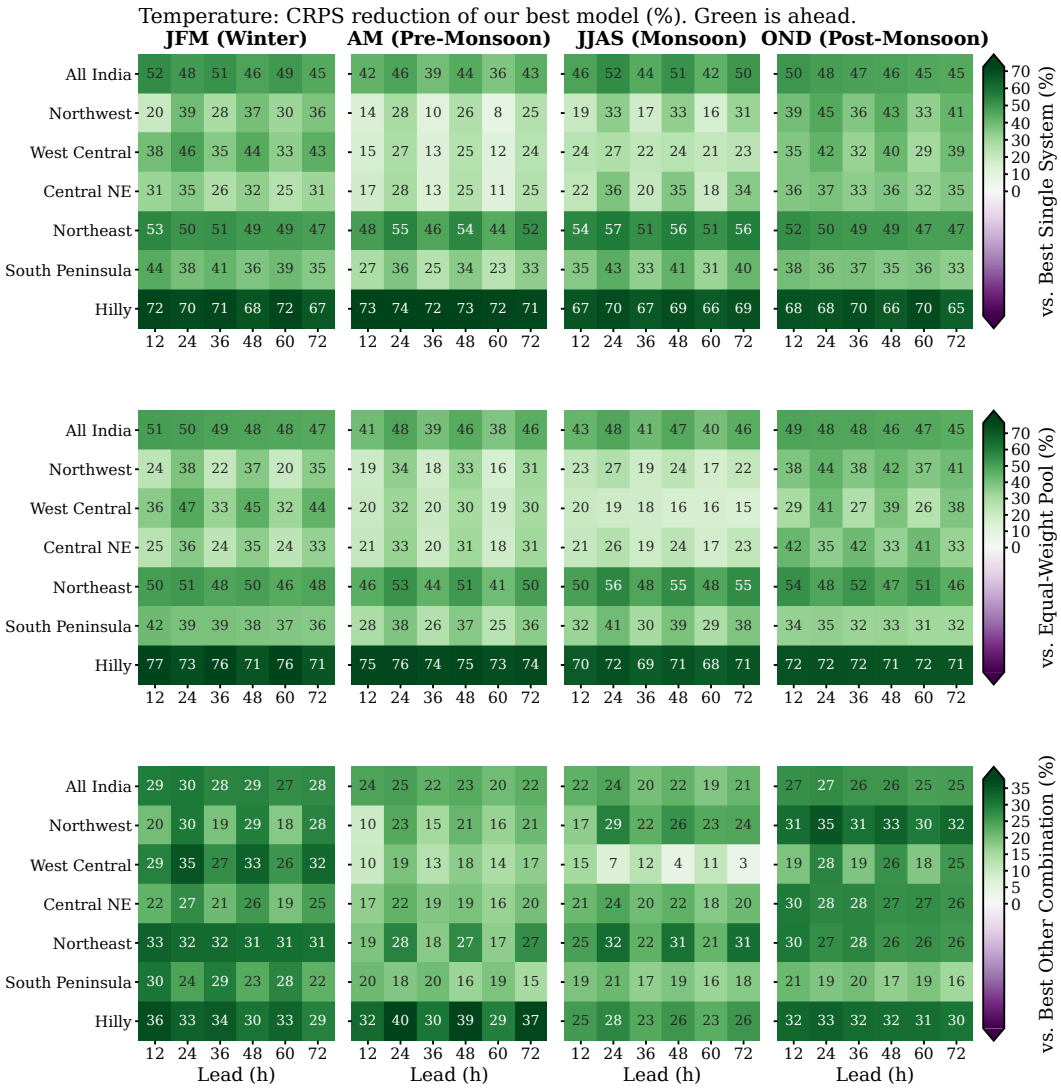


Figure 5: **Temperature: CRPS reduction of our best model (%),** by lead (columns within a block), IMD homogeneous region (rows) and season (blocks). Green is ahead.

the challenging 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. Figure 3 (a) (b) show our method having lower regret than others for temperature, with similar results for precipitation shown in Figure 6. Fig 3 (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 Appendix C. Figure 4 visually shows improvements relative to other models, with significant improvements visible in North India and central Europe.

Where the gains come from We can also break the reduction down by lead, region, season and comparator simultaneously, as shown in Figure 5.

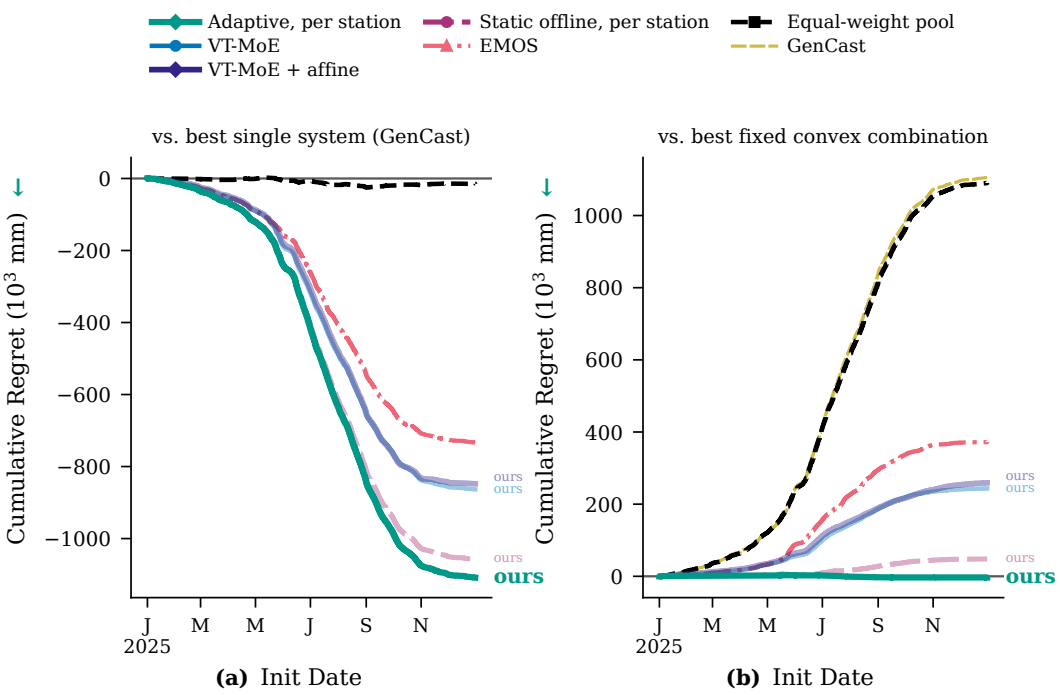


Figure 6: **Cumulative regret on rainfall,** against (a) the best single system and (b) the best fixed convex combination, both chosen in hindsight over the same window.

5 Conclusion

We show a framework that learns a forecast from historical forecast-observation pairs, followed by online adaptation with an algorithm that has formal relative error bounds on the combination method that 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 (Brown et al., 2025) and Tessera (Feng et al., 2026), 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.

AI use statement

We have used generative AI and LLMs for literature review and contextualising our work, for polishing the grammatical and writing structure of the paper, and assisting in generation of the framework figure, validating the proofs and debugging the code.

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.

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.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicoló 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. ISSN 0004-5411. doi: 10.1145/258128.258179.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In Peter L. Bartlett, Fernando Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 989–997. Curran Associates, Inc., 2012.

Kai Chang and Themistoklis P Sapsis. Extreme event aware (η-) learning. *Nature Communications*, 2026.

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.

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*, pp. 34818–34831, 2026.

Yoav Freund. Predicting a binary sequence almost as well as the optimal biased coin. Technical report, AT&T Research, 1996.

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.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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.

Hakan Gokcesu and Suleyman S. Kozat. Optimal tracking in prediction with expert advice. *arXiv preprint arXiv:2208.03708*, 2022. doi: 10.48550/arXiv.2208.03708.

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.

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.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

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.

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.

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.

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.

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

Vladimir G. Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pp. 51–60. ACM, 1995.

Volodimir G Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pp. 371–386. Morgan Kaufmann, 1990.

Vladimir V. V'yugin and Vladimir Trunov. Online aggregation of probability forecasts with confidence. *arXiv preprint arXiv:2109.14309*, 2021.

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

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**
Forecast experts			
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.
Station observations			
Domain	**Network**	**Time period**	**Stations**
India	IMD AWS	2019–2025	1296
Europe	Stationbench	2022–2025	3672
CONUS	Stationbench	2022–2025	3032

AWS refers to the Automated Weather Station Network of the IMD. This is 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. The count of the number of stations are after filtering for those with at least 200 verifications in the window. The following spatial domains are used: India 6–38°N, 68–98°E; Europe 35–72°N, 25°W–45°E; CONUS 24–50°N, 125–66.5°W.

The temperature variable is taken from all the models, and precipitation from Gencast, GEFS, and GEM.

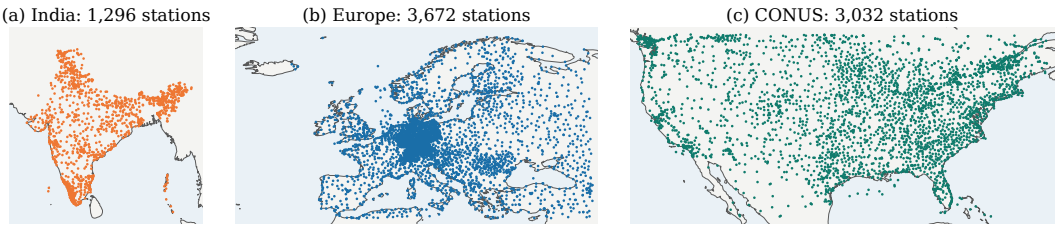


Figure 7: **The three verification networks.**

Why combine at all. The premise of the whole framework is that no single system is reliably best. Figure 8 resolves that station by station: on temperature FGN wins at 44 % of Indian stations and GEM at 22 %, with the rest split three ways, and on rainfall GEFS and GenCast divide up the rest of the stations almost equally.

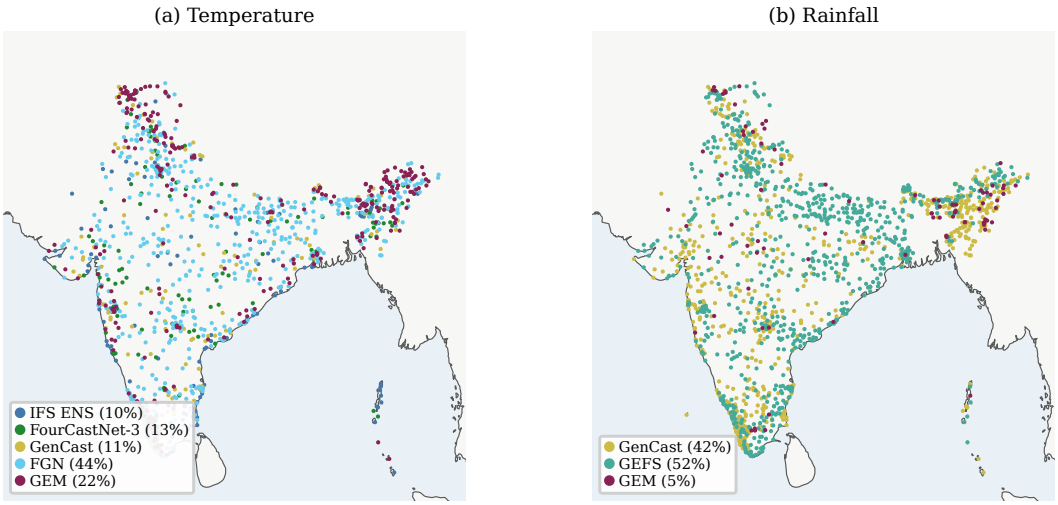


Figure 8: **No single system is best everywhere.**

B Methods & Experiments

Setup. Index forecast issue times by $t \in [T]$, lead times by $\tau \in \mathcal{T} = \{12, \ldots, 72\}$ 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 $t + \tau$.

Offline Supervised Combiner: Architecture and Training Our offline aggregator is a U-Net that maps expert forecasts to a probabilistic mixture forecast over the target ground truth at L lead times. It learns which forecast models to weigh more based on historical data. The encoder applies four `DoubleConv` stages of channels $(32, 64, 128, 256)$ separated by 2×2 max-pool downsamples, and the decoder mirrors it with transposed-conv upsamples and U-Net skip connections from the encoder. Two 1×1 conv heads on the final feature map produce expert logits and a scalar bias map. A softmax over experts gives per-pixel convex weights over the station grids

$$w_{i,n,h,w} = \frac{\exp(z_{i,n,h,w})\, m_{i,n}}{\sum_{n'} \exp(z_{i,n',h,w})\, m_{i,n'}}, \qquad \sum_{n} w_{i,n,h,w} = 1,$$

where $m_{i,n} \in \{0, 1\}$ is a per-init expert-availability mask that excludes experts with missing data on a given init time.

Training Each expert's per-pixel mean $\mu_n \in \mathbb{R}^{H \times W}$ and standard deviation $\sigma_n \in \mathbb{R}^{H \times W}$ are stacked into the input $S_t \in \mathbb{R}^{N \times 2 \times H \times W}$. The U-Net takes (S_t, ℓ) and produces per-pixel weights $w \in \mathbb{R}^{N \times H \times W}$, which induce a combined predictive CDF $\hat{F}_t^{(w)}$. We train end-to-end against

$$L_\theta(S_t, \ell) = \mathrm{CRPS}\big(\hat{F}_t^{(w)}, y_\ell\big),$$

where $y_\ell \in \mathbb{R}^{H \times W}$ is the ground truth at lead time ℓ.

We train θ to minimise the fair (kernel) CRPS of the combined forecast against the station ground truth y.

Optimisation and runtime. We optimise using AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 8) for 30 epochs, with a cosine-annealing schedule down to $\eta_{\min} = 10^{-6}$. Each training example is an (init, lead) pair, so the model sees all $L = 6$ leads independently and the FiLM head learns lead-specific behaviour. Years 2019–2023 are used for training, and 2024–2025 are held out for testing. End-to-end training takes ≈ 5 h on a single H100. All inputs and ground-truth tensors are stored as Zarr at 0.25° resolution and accessed via xarray.

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 $m_r = \bar{q}_r(0.5)$:

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

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 (Cesa-Bianchi & Lugosi, 2006), $\eta = \min\{1/L_{\max}, 1.0739\sqrt{\ln N/V}\}$ 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.

VT-MoE. More details in appendix E.

EMOS Ensemble model output statistics (Gneiting et al., 2005) is the post-processing usually used operationally so we take it as a baseline. We fit the Gaussian form

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

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 and not by maximum likelihood, using the closed form for a Gaussian,

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

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 Empirical CRPS For a finite ensemble $\{x_1, \ldots, x_M\} \subset \mathbb{R}$ sampled from a predictive distribution F, we use the unbiased $M(M-1)$-normalised plug-in estimator

$$\widehat{\mathrm{CRPS}}(\{x_i\}, y) \;=\; \underbrace{\frac{1}{M}\sum_{i=1}^{M} |x_i - y|}_{\text{skill term}} \;-\; \underbrace{\frac{1}{M(M-1)} \sum_{1 \le i < j \le M} \big(x_{(j)} - x_{(i)}\big)}_{\text{spread term}}, \tag{10}$$

where $x_{(1)} \le x_{(2)} \le \cdots \le x_{(M)}$ denote the order statistics of the ensemble. The closed-form

$$\mathrm{CRPS}(F, y) \;=\; \tfrac{1}{M}\sum_{i} |e_i - y| \;-\; \tfrac{1}{2M^2}\sum_{i,j} |e_i - e_j|$$

is referenced in the main text.

C Europe, USA Results

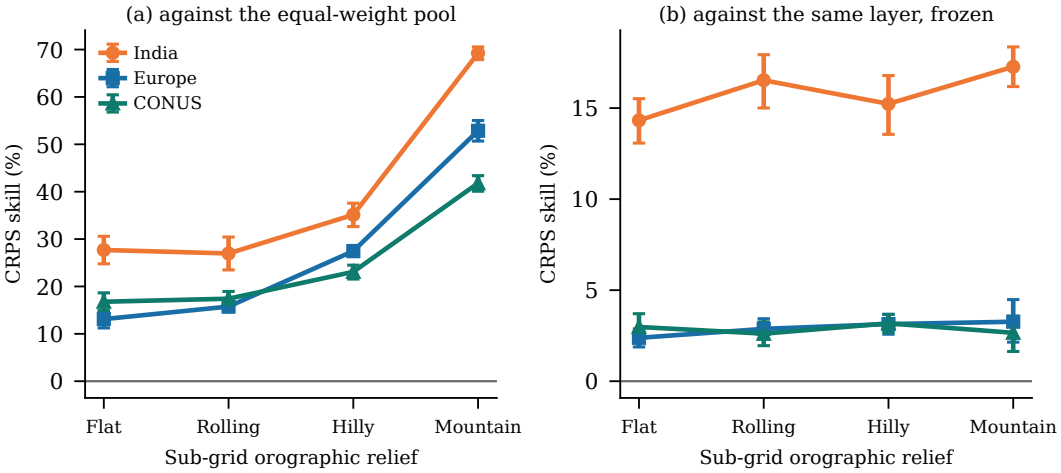
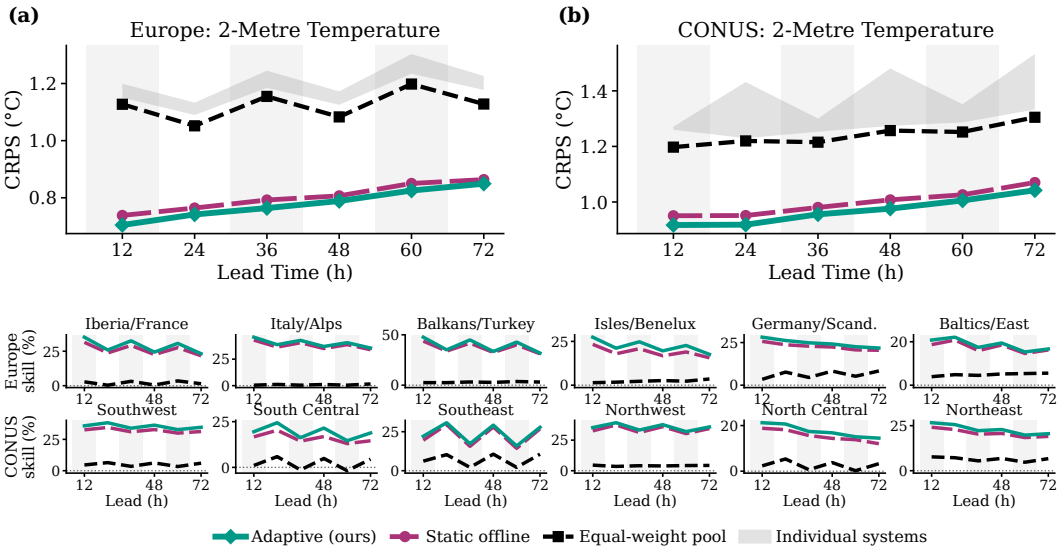


Figure 9: **Skill against sub-grid orographic relief, three networks.** Higher is better. (a) The gain over an equal-weight pool rises steeply with relief on every network (b) The gain from adapting online rather than freezing the same layer is far flatter, and several times larger in India which shows 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 10: **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 (°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 ($1.037 \to 0.906$) while adapting only the weights recovers markedly less ($1.037 \to 0.965$).

Why the affine term beats VT-MoE, and why VT-MoE is still worth it 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.

VTMOE having worse empirical performance than our best method 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, and what it cannot do is leave this class. These two are complementary since 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.

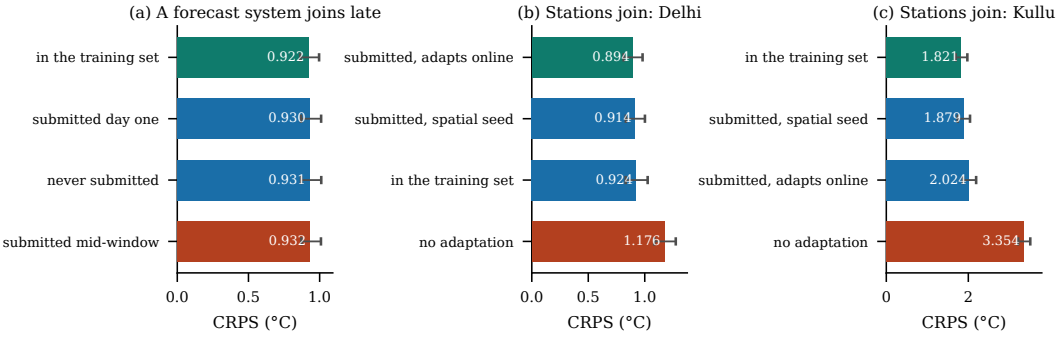


Figure 11: **What submitting data buys.** Lower is better (a) A forecast (FGN) is withheld from the panel and then submitted part-way through. Scored on the rows after the submission date, arriving mid-window is worth as much as having been present from the first day, and both sit inside the interval 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 of 74 stations near Delhi on the plain and 65 near Kullu in the western Himalaya. Here the picture is quite different and we see 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 Proof of Theorem 1

In this section, our aim is to prove Theorem 1. First, we begin with a literature review on the field of expert advice, motivating the problem addressed by us. Subsequently, we prove the theorem assuming a series of intermediate results, which are then established in the subsequent section.

E.1 Literature Review and Motivation

The field of expert advice can be traced back to the articles of Vovk(Vovk, 1990; 1995). Multiple influential papers in the field include those by Cesa-Bianchi (Cesa-Bianchi et al., 1997; 2012), Haussler (Haussler et al., 1994) and Freund (Freund, 1996). More recent articles on the subject include (Gokcesu & Kozat, 2022). We refer the reader to (Cesa-Bianchi & Lugosi, 2006) for standard literature and reading material on the subject.

Formally, fix a number $T \in \mathbb{N}$ and outcome space Y. Suppose that $y_t, t \in [T]$ is a sequence of unknown observations, which we assume to be adversarial (i.e. carrying no inherent statistical or deterministic pattern). At each $t \in [T]$, N experts attempt to predict the unknown element y_t; their predictions lie in a decision space D. Let $D_{i,t}$ The learner receives the expert predictions for the round, and combines them using some algorithm to make a prediction $\hat{D}_t$. Finally, there is a non-negative loss function $\ell : D \times Y \to [0, \infty)$ which penalizes poor predictions.

The aim of the learner is to design a combination algorithm that produces low regret. Formally, let

$$L_{i,T} = \sum_{t=1}^{T} \ell(D_{i,t}, y_t), \quad L_T = \sum_{t=1}^{T} l(\hat{D}_t, y_t).$$

Furthermore, if D is a convex set, then one can consider static convex combinations of expert predictions. If Θ is as in (2) and $(p_1, \dots, p_n) \in \Theta$, then we can define

$$L_{p,T} = \sum_{t=1}^{T} \ell\left(\sum_{i=1}^{N} p_i D_{i,t}, y_t\right).$$

The performance of the learner is typically evaluated using the *regret*. For instance, consider the regret against the *best expert*

$$R'_T = L_T - \min_{i \in [N]} L_{i,T}. \tag{11}$$

However, comparing oneself against the best expert is insufficient if one intends to beat each expert. For instance, there is no benefit of combining weather models, if the combination thus produced does not forecast better than the individual models themselves. With this in mind, the notion of regret against the best *mixture of experts*, given by

$$R_T = L_T - \min_{p \in \Theta} L_{p,T}. \tag{12}$$

It is in the stronger setup that Theorem 3 is established.

E.2 Auxiliary results, and proof of Theorem 1

Now, we shall state some auxiliary results required for the proof of Theorem 1. Subsequently, we prove the theorem assuming these results.

An important role in the proof is played by the notion of mixability. For mixable loss functions, one can easily design an algorithm which admits strong regret guarantees. While a formal definition can be found in (V'yugin & Trunov, 2021), we state the precise definition for the mixture-of-experts regret setup.

Definition 2. Fix an outcome space Y, a convex decision space D and $N \in \mathbb{N}$. A non-negative loss function $\ell : D \times Y \to [0, \infty)$ is called η-mixable for some $\eta > 0$ if, for every collection of expert predictions $\{F_n\}_{n=1}^{N} \subset D$ and every finite measure ν on Θ, there exists a prediction $F \in D$ such that for all $y \in Y$,

$$\ell(F, y) \;\leq\; -\frac{1}{\eta} \ln \frac{\int_\Theta \exp\Big(-\eta\, \ell\left(\sum_{n=1}^{N} p_n F_n, y\right)\Big)\, d\nu(p_1, \dots, p_N)}{\nu(\Theta)}. \tag{13}$$

Our first auxiliary result is of independent interest; we establish the mixability of the CRPS loss as per the definition above.

Theorem 3 (Mixability)**.** *The CRPS loss is $\frac{2}{b-a}$-mixable in the MOE setting. Moreover, inequality* (13) *is satisfied by the explicit CDF*

$$\hat{F}(y) \;=\; \tfrac{1}{2} \;-\; \tfrac{1}{4}\,\ln\!\left(\frac{\int_\Theta \exp\!\big(-2\big[F^{(p)}(y)\big]^2\big)\,d\nu(p)}{\int_\Theta \exp\!\big(-2\big[1-F^{(p)}(y)\big]^2\big)\,d\nu(p)}\right). \tag{14}$$

Next, we state some desirable properties of the CRPS loss, that explain both the advantage of our bounded observation setup, and the quadratic nature of the CRPS loss evaluated at mixtures.

Lemma 4 (Quadratic structure)**.** *Let $L(p) = \sum_{t=1}^{T} \mathrm{CRPS}(F_t^{(p)}, y_t)$, and define $A_i = \sum_t \mathrm{CRPS}(F_{i,t}, y_t)$ 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 $1 \le i, j \le N$ and $p \in \Theta$:* (i) $|A_i - A_j| \le 2(b-a)T$*;* (ii) $|(Bp)_i - (Bp)_j| \le (b-a)T$*;* (iii) $-\delta^\top B\delta \le (b-a)NT\|\delta\|_2^2$ *for every δ with $\mathbf{1}^\top\delta = 0$.*

Lemma 6. *For a sequence of observations $y_t, t \in [T]$ and expert predictions $F_{i,t}, i \in [N], t \in [T]$, let $\hat{F}_t, t \in [T]$ be the sequence of predictions in Algorithm 1, and for some $\mathsf{M} > 0$ let $\hat{F}_{\mathsf{M},t}, t \in [T]$ be the sequence of random predictions in Algorithm 2. Let*

$$K = \sum_{t=1}^{T} \int_a^b (\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2 dx.$$

Then, for any $\delta > 0$,

$$\mathbb{P}(K > \delta) \le \frac{4(b-a)}{\delta \mathsf{M}} \sum_{t=1}^{T} e^{4t}(1+4t^2).$$

We are ready to prove our result.

Proof of Theorem 1. We begin with the proof of (7) and (9), which are the regret bounds for Algorithm 1. Note that as $M \to \infty$, by the law of large numbers, the left hand side of each of the above equations is precisely the regret of Algorithm 7.

Our first step is to bound the loss of Algorithm 2. Note that it maintains a measure μ_t on Θ, initialized to the uniform prior μ_0 (a choice supported by the two-expert analysis of Freund, 1996, Sections 4,7). At round t it outputs a CDF $\hat{F}_t$ satisfying (13) relative to μ_t, then updates

$$\frac{d\mu_{t+1}}{d\mu_t}(p) \;=\; \exp\!\big(-\eta\;\mathrm{CRPS}(F_t^{(p)}, y_t)\big). \tag{15}$$

Theorem 3 establishes that such an $\hat{F}_t$ exists in closed form, with $\eta = 2/(b-a)$. Iterating (13) over T rounds shows that

$$\sum_{t=1}^{T} \mathrm{CRPS}(\hat{F}_t, y_t) \le -\ln \mu_{T+1}(\Theta). \tag{16}$$

Substituting the explicit form of $\hat{F}_t$ (Theorem 3) into (15) and unrolling, the prior μ_t collapses into the predicted integrand.

Let $p^* \in \arg\min_\Theta \sum_{t=1}^{T} \mathrm{CRPS}\left(\sum_{n=1}^{N} p_n F_{n,t}, y_t\right)$, which exists since a continuous function over a compact set always attains its minimum. Iterating the equation (15) over $t \in [T]$ and applying Lemma 4 we obtain

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

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 $p = (p_1, \ldots, p_N)$. Now, by (16) and Lemma 4,

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

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

$$\mathrm{Reg}_T \leq -\frac{b-a}{2} \ln \int_\Theta e^{-\frac{2}{b-a}(L(p_1,\ldots,p_N) - L(p^*))} dp. \tag{17}$$

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 $\nabla L(p^*) = A - Bp^*$, 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

$$L(p) - L(p^*) = (A - Bp^*)^T \delta - \frac{1}{2} \delta^T B \delta. \tag{18}$$

Observe that we do not assume anything about the point p^* at all during this manipulation: indeed, if p^* is interior to Θ then $(A - Bp^*) = 0$ 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 shall first prove the bound (7), assuming no structure on p^*. Then, we shall specialize to (9).

We bound the right hand side of (18), beginning with the first term. By Lemma 5,

$$\begin{aligned}
(A - Bp^*)^T \delta =& (A - Bp^* + \min_i (A - Bp^*)_i \mathbf{1})^T \delta \\
\leq& \max_{1 \leq i,j \leq N} |(A_i - (Bp^*)_i) - (A_j - (Bp^*)_j)| \times \|\delta\|_1 \\
\leq& \left[\max_{1 \leq i,j \leq N} |A_i - A_j| + \max_{1 \leq i,j \leq n} |(Bp^*)_i - (Bp^*)_j|\right] \times \|\delta\|_1 \\
\leq& [2(b-a)T + (b-a)T] \times \|\delta\|_1 \\
\leq& 3(b-a)T\|\delta\|_1 \leq 3(b-a)T\sqrt{N}\|\delta\|_2,
\end{aligned} \tag{19}$$

where we used the Cauchy-Schwarz inequality in the last line.

For the other term, we simply have

$$-\frac{1}{2}\delta^T B \delta \leq (b-a)NT\|\delta\|_2^2 \tag{20}$$

by Lemma 5. Combining the above two bounds and (18),

$$L(p) - L(p^*) \leq 3(b-a)T\sqrt{N}\|\delta\|_2 + (b-a)NT\|\delta\|_2^2 \leq (b-a)T(3\sqrt{N} + N)\|\delta\|_2 \tag{21}$$

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 (21) implies that $L(p) - L(p^*) \leq b - a$. In particular,

$$\|p^* - p\| \leq r \implies e^{-\frac{2}{b-a}(L(p) - L(p^*))} \geq e^{-2}. \tag{22}$$

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

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

where $|S|$ is the $(N-1)$-dimensional Lebesgue measure of S. This combined with (17) yields

$$\mathrm{Reg}_{M,T} \le -\frac{b-a}{2}\ln e^{-2}|S| \le \frac{b-a}{2}\ln\frac{1}{|S|} + (b-a). \tag{23}$$

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

$$\ln\frac{1}{|S|} \le \ln\frac{1}{c_\Theta|\mathbf{B}(p^*,r)|} = (N-1)\ln\frac{1}{r} + C = (N-1)\ln(T) + C \tag{24}$$

where C is independent of T. Finally, combining this with (23) gives

$$\mathrm{Reg}_T \le \frac{(b-a)(N-1)}{2}\ln(T) + C$$

where C has no dependence on T, completing the proof of (7).

For the proof of (9), we note that if p^* is an interior minimum, then the bound on the term (19) is not required since $(A - Bp^*) = 0$. Hence, using merely the bound (20),

$$L(p) - L(p^*) \le (b-a)NT\|\delta\|_2^2.$$

Now, if we take $r = \frac{1}{\sqrt{NT}}$, then $\|\delta\|_2 < r$ implies that $L(p) - L(p^*) \le (b-a)$. Thus, repeating the argument from (22) to (24) gives

$$\ln\frac{1}{|S|} \le \ln\frac{1}{c_\Theta|\mathbf{B}(p^*,r)|} = (N-1)\ln\frac{1}{r} + C = \frac{(N-1)}{2}\ln(T) + C$$

whence continuing the proof as earlier leads to the bound (9), as desired. We have proved the regret bounds for Algorithm 1.

Now, we shall prove the regret bounds for Algorithm 2, namely (6) and (8). We shall only prove (6) as a consequence of (7). The proof of (8) as a consequence of (9) is identical.

For this, we begin on a relatively abstract note, with the aim to connect (6) and (8). Let $L^2[a,b])$ be the space of square integrable functions on $[a,b]$ with respect to the Lebesgue measure. Note that its T-fold product $(L^2[a,b])^T$ can be equipped with the usual norm

$$\|(f_1, f_2, \ldots, f_t)\|^2_{L^2([a,b])^T} = \sum_{t=1}^{T}\int_a^b f_i(x)^2 dx. \tag{25}$$

Consider any sequence of CDFs $F_1, F_2, \ldots, F_T$ of random variables concentrated on $[a,b]$, and outcomes $y_1, y_2, \ldots, y_T \in [a,b]$. The sequence of CDFs can be written as a vector of the form $(F_1, F_2, \ldots, F_T) \in (L^2[a,b])^T$. Similarly, if we consider each outcome y_i as a constant random variable, its CDF is given by $x \mapsto 1_{y_i \le x}$, whence $(y_1, \ldots, y_T)$ also embeds itself into $(L^2[a,b])^T$. By the definition of the CRPS distance (1),

$$\sum_{t=1}^{T}\mathrm{CRPS}(F_T, y_t) = \|(F_1, F_2, \ldots, F_T) - (x \mapsto 1_{y_1 \le x}, x \mapsto 1_{y_2 \le x}, \ldots, x \mapsto 1_{y_T \le x})\|^2_{(L^2[a,b])^T} \tag{26}$$

for any $(F_1, \ldots, F_T) \in (L^2([a,b]))^T$. Now, we will apply the following inequality which is easily verified: for any four vectors a, b, c, d in a normed space with norm $\|\cdot\|$, if $\|b - c\| \le R$ for some $R > 0$, then we have

$$\sqrt{\|a-c\|^2 - \|b-c\|^2} \le \sqrt{\|d-c\|^2 - \|b-c\|^2} + \sqrt{2R\|d-a\| + \|d-a\|^2}. \tag{27}$$

We apply the above inequality with carefully chosen vectors in the space $L^2([a,b])^T$ with the norm (25). Given a sequence of outcomes $y_1, y_2, \ldots, y_T \in [a,b]$ and expert predictions $F_{i,t}, i \in [N], t \in [T]$, let $\hat{F}_t, \hat{F}_{\mathsf{M},t}$ be the predictions at time $t \in T$ given by Algorithms 1 and 2 respectively, for some $\mathsf{M} > 0$. Let $L(p) = \sum_{t=1}^{T} \mathrm{CRPS}(\sum_{n=1}^{N} p_n F_{n,t}, y_t)$ and $p^* \in \arg\min L(p)$ be as in the previous proof.

We apply the inequality (27) with the following choice of sequences: $a = (\hat{F}_t, t \in [T])$, $b = (\sum_{n=1}^{N} p_n^* F_{n,t}, t \in [T])$, $c = (y_t, t \in [T])$ (note: this corresponds naturally to a sequence of CDFs as mentioned earlier), and $d = (\hat{F}_{\mathsf{M},t}, t \in [T])$. Observe that $\|b - c\|^2 = L(p^*)$. By (26) and the definition (4) of the regret,

$$\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \sqrt{2\sqrt{L(p^*)K} + K}, \tag{28}$$

where

$$K = \sum_{t=1}^{T} \int_a^b (\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2 dx$$

is as in Lemma 6. Note that K is the only probabilistic quantity in (28). Under perfect integration, we would have had $K = 0$, leading to equality. However, if M is large, Lemma 6 derives concentration bounds on K which allows us to complete the result as shown below.

The following series of implications determine the precise constant for which Lemma 6 needs to be applied. Let $\epsilon > 0$ be arbitrary. Note that

$$K \leq \epsilon^2 + 2L(p^*) - 2\sqrt{L(p^*)^2 + L(p^*)\epsilon^2} \implies \sqrt{2\sqrt{L(p^*)K} + K} \leq \epsilon.$$

As a consequence of the above inequality and (28),

$$\mathbb{P}\left(K \leq \epsilon^2 + 2L(p^*) - 2\sqrt{L(p^*)^2 + L(p^*)\epsilon^2}\right) \leq \mathbb{P}\left(\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \epsilon\right).$$

Now, applying Lemma 6,

$$\mathbb{P}\left(\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \epsilon\right) \geq 1 - \frac{4(b-a)}{\mathsf{M}(\epsilon^2 + 2L(p^*) - 2\sqrt{L(p^*)^2 + L(p^*)\epsilon^2})} \sum_{t=1}^{T} e^{4t}(1 + 4t^2).$$

Finally, applying the first part of this theorem,

$$\mathbb{P}\left(\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \epsilon\right) \geq 1 - \frac{4(b-a)}{\mathsf{M}(\epsilon^2 + 2L(p^*) - 2\sqrt{L(p^*)^2 + L(p^*)\epsilon^2})} \sum_{t=1}^{T} e^{4t}(1 + 4t^2).$$

In order to make the right hand side independent of $L(p^*)$, we note that $L(p^*) \leq T(b-a)$ by definition, and

$$x \mapsto \epsilon^2 + 2x - 2\sqrt{x^2 + x\epsilon^2}$$

is a monotone decreasing mapping on the domain $x \in (0, \infty)$. As a consequence,

$$\mathbb{P}\left(\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \epsilon\right) \geq 1 - \frac{4(b-a)}{\mathsf{M}(\epsilon^2 + 2T(b-a) - 2\sqrt{T^2(b-a)^2 + T(b-a)\epsilon^2})} \sum_{t=1}^{T} e^{4t}(1+4t^2). \tag{29}$$

To end the proof, we perform some routine algebraic and analytic simplifications on the right hand side. First, observe the following two identities:

$$\lim_{x \to 0} \frac{\epsilon^2 + 2x - 2\sqrt{x^2 + x\epsilon^2}}{\epsilon^2} = 1,$$

and

$$\lim_{x \to \infty} \frac{\epsilon^2 + 2x - 2\sqrt{x^2 + x\epsilon^2}}{\frac{\epsilon^4}{4x}} = 1.$$

Combining the two equations above, there exists a universal constant C such that

$$\epsilon^2 + 2x - 2\sqrt{x^2 + x\epsilon^2} \geq C \min\left\{\epsilon^2, \frac{\epsilon^4}{4x}\right\}.$$

Plugging the above into (29),

$$\mathbb{P}\left(\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \epsilon\right) \geq 1 - \frac{4C(b-a)}{\mathsf{M}\min\{\epsilon^2, \frac{\epsilon^4}{4T(b-a)}\}} \sum_{t=1}^{T} e^{4t}(1+4t^2). \tag{30}$$

To bound the term in T, it is sufficient to note that

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} e^{4t}(1+4t^2)}{\frac{4e^4}{e^4-1}T^2e^{4T}} = 1.$$

As a consequence, for some constant C independent of T, we have

$$\sum_{t=1}^{T} e^{4t}(1+4t^2) \leq CT^2e^{4T}.$$

Combining this with (30) and absorbing the $4C(b-a)$ into the minimum,

$$\mathbb{P}\left(\sqrt{\mathrm{Reg}_{\mathsf{M},T}} \leq \sqrt{\mathrm{Reg}_T} + \epsilon\right) \geq 1 - \frac{4C(b-a)}{\mathsf{M}\min\{\epsilon^2, \frac{\epsilon^4}{4T(b-a)}\}} T^2e^{4T}.$$

This completes the proof. □

E.3 Proofs of Auxiliary results, and minimax optimality

In this section, we will prove the auxiliary results used in the proof of Theorem 1. We prove them in the order they were stated. We end the section with a detailed discussion on the minimax optimality of Algorithm 1.

We begin with the proof of Theorem 3. In order to do this, we require the following result which is informally stated in (Freund, 1996), and is a continuous version of the analogous result in (Haussler et al., 1994). The proof is a direct calculation which we will omit.

Proposition 7. *Consider the loss function $l' : \{0,1\} \times [0,1] \to [0,1]$ 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 3. We shall borrow several notations and arguments from the proof of (V'yugin & Trunov, 2021, 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 $k = 0, \ldots, d$. 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

$$F_d(z) = \begin{cases} 0 & z \leq a \\ 1 & z \geq b \\ F(z_k) & z \in [z_{k-1}, z_k], 1 \leq k \leq d. \end{cases} \tag{31}$$

One may consider F_d as a "discretization" of F. First, we can derive the bound below exactly as in (V'yugin & Trunov, 2021, equation (21)).

$$|\operatorname{CRPS}(F_d, y) - \operatorname{CRPS}(F, y)| \leq 2\Delta \tag{32}$$

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

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

Once again, exactly as in (V'yugin & Trunov, 2021, equation (22)), we obtain

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

for all $y \in \mathbb{R}$. Combining this with (32) we have

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

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

$$\lim_{d \to \infty} \Delta \sum_{s=1}^{d} (F(z_s) - \omega_{y,s})^2 = \operatorname{CRPS}(F, y). \tag{33}$$

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

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

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

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

$$\pi_s(A) = \frac{\int_\Theta 1_{\{\sum_{n=1}^N p_n F_n(z_s) \in A\}} d\nu(p)}{\nu(\Theta)}. \tag{34}$$

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

$$e^{-2l'(\hat{f}_s, \Omega_s)} \geq \int_0^1 e^{-2l'(t, \Omega_s)} d\pi_s(t) \tag{35}$$

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 7. By the definition (34) of π_s and Fubini's theorem we have

$$\begin{aligned} \int_0^1 e^{-2l'(t, \Omega_s)} d\pi_s(t) &= \frac{\int_\Theta e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \\ \hat{f}_s &= \frac{1}{2} - \frac{1}{4} \ln \left(\frac{\int_\Theta e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), 0)} d\nu(p)}{\int_\Theta e^{-2l'(\sum_{n=1}^N p_n F_n(z_s), 1)} d\nu(p)} \right). \end{aligned} \tag{36}$$

Combining the first of these with (35),

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

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

$$e^{-2\Delta \sum_{s=1}^{d} l'(\hat{f}_s, \Omega_s)} \geq \prod_{s=1}^{d} \left(\frac{\int_\Theta e^{-2l'(\sum_{n=1}^{N} p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)} \right)^{\Delta}. \tag{37}$$

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 (37) 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

$$e^{-\frac{2}{b-a} \Delta \sum_{s=1}^{d} l'(\hat{f}_s, \Omega_s)} \geq \frac{\int_\Theta e^{-\frac{2}{b-a} \Delta \sum_{s=1}^{d} l'(\sum_{n=1}^{N} p_n F_n(z_s), \Omega_s)} d\nu(p)}{\nu(\Theta)}. \tag{38}$$

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

$$e^{-\frac{2}{b-a} \Delta \sum_{s=1}^{d} (\hat{f}_s - \omega_{y,s})^2} \geq \frac{\int_\Theta e^{-\frac{2}{b-a} \Delta \sum_{s=1}^{d} l'(\sum_{n=1}^{N} p_n F_n(z_s), \omega_{y,s})} d\nu(p)}{\nu(\Theta)}, \tag{39}$$

where by (36) 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 $d \to \infty$ in (39), by Fubini's theorem and (33) we obtain

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

This completes the proof. □

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

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

On the other hand,

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

Subtracting the previous equation from this one,

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

However, we note the algebraic identity

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

Combining this with the previous equality,

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

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. □

Next, we prove Lemma 5.

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

$$\int_a^b (F(x) - G(x))^2 dx \le \int_a^b 1 dx \le (b - a). \tag{40}$$

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

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

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

since $p_i, p_j \in [0, 1]$ and B has non-negative entries. However, this quantity is bounded by $T(b - a)$ by (40). The result follows.

Finally, for part (3) let $\delta = (\delta_1, \ldots, \delta_N)$ 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 $\left(\sum_{n=1}^{N} \delta_n F_{n,t}(z)\right)^2 \leq \left(\sum_{n=1}^{N} \delta_n)\right)^2 \leq \|\delta\|_2^2$ by the Cauchy-Schwarz inequality. Applying this to the previous inequality, we obtain

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

Thus, the proofs of all three parts are complete. □

Finally, we prove Lemma 6.

Proof of Lemma 6. Let $\hat{F}_t, \hat{F}_{\mathsf{M},t}, t \in [T]$ and K be as in the statement of the lemma. Let $\delta > 0$ be arbitrary.

We start with the Markov inequality

$$\mathbb{P}(K > \delta) \leq \frac{\mathbb{E}\left[\sum_{t=1}^{T} \int_a^b (\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2 dx\right]}{\delta} = \frac{\sum_{t=1}^{T} \int_a^b \mathbb{E}\left[(\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2\right] dx}{\delta}, \quad (41)$$

where we used the Fubini-Tonelli theorem to interchange the integrals, sums and expectations in the numerator.

Now, we will upper bound the inner term carefully, for fixed $x \in [a,b], t \in [T]$. We introduce some convenient notation now. Recall Θ from (2) and let $g_0, g_1, w_t : \Theta \to \mathbb{R}$ be defined by

$$g_0(p) = \exp(-2[F_t^p(x)]^2), \quad g_1(p) = \exp(-2[1 - F_t^p(x)]^2), \text{ and} \quad (42)$$

$$w_t(p) = \exp\left(-\frac{2}{b-a} S_{t-1}(p)\right). \quad (43)$$

Now, from the definitions of $\hat{F}_t$ and $\hat{F}_{\mathsf{M},t}$ from Algorithms 1 and 2 respectively,

$$\hat{F}_t(x) = \frac{1}{2} - \frac{1}{4} \log \frac{\mathbb{E}_U[g_0 w_t]}{\mathbb{E}_U[g_1 w_t]},$$

$$\hat{F}_{\mathsf{M},t}(x) = \frac{1}{2} - \frac{1}{4} \log \frac{\frac{1}{\mathsf{M}} \sum_{j=1}^{\mathsf{M}} g_0(v_j) w_t(v_j)}{\frac{1}{\mathsf{M}} \sum_{j=1}^{\mathsf{M}} g_1(v_j) w_t(v_j)},$$

where $v_j = (v_{1j}, v_{2j}, \ldots, v_{Nj})$ are i.i.d. uniformly chosen samples from Θ as per Algorithm 2.

Some more convenient notation follows. For $i \in \{0,1\}$, let

$$A_i = \mathbb{E}_U[g_i w_t], \quad \hat{A}_i = \frac{1}{\mathsf{M}} \sum_{j=1}^{\mathsf{M}} g_0(v_j) w_t(v_j). \quad (44)$$

We have

$$\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x) = \frac{1}{4}\left(\log \frac{\hat{A}_0}{\hat{A}_1} - \log \frac{A_0}{A_1}\right) = \frac{1}{4}\left(\log \frac{\hat{A}_0}{A_0} + \log \frac{\hat{A}_1}{A_1}\right).$$

The following inequality is now clear:

$$\mathbb{E}\left[(\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2\right] = \frac{1}{16}\mathbb{E}\left[\left(\log \frac{\hat{A}_0}{A_0} + \log \frac{\hat{A}_1}{A_1}\right)^2\right] \leq \frac{1}{8}\left(\mathbb{E}\left[\log^2 \frac{\hat{A}_0}{A_0}\right] + \mathbb{E}\left[\log^2 \frac{\hat{A}_1}{A_1}\right]\right). \quad (45)$$

Thus, we have reduced the problem to bounding the expected logarithm squared of two empirical quantities, which we proceed to do right now.

Consider the event $G_i = \left\{\frac{\hat{A}_i}{A_i} \in \left[\frac{1}{2}, \frac{3}{2}\right]\right\}$ for $i \in \{0,1\}$. On this event, since $|\log(x)| \leq 2|x-1|$ for $x \in [\frac{1}{2}, \frac{3}{2}]$ and $\mathbb{E}[\hat{A}_i] = A_i$, we obtain that

$$\mathbb{E}\left[\log^2 \frac{\hat{A}_i}{A_i} 1_{G_i}\right] \leq \frac{4}{A_i^2} \mathrm{Var}(A_i) = \frac{4}{\mathsf{M}} \frac{Var_U(g_i w_t)}{\mathbb{E}_U[g_i w_t]^2},$$

where we recall g_0, g_1, w_t from (42) and (43). We now use the fact that $e^{-2} \le g_i \le 1$ for $i = 0, 1$ to get

$$\mathbb{E}\left[\log^2 \frac{\hat{A}_i}{A_i} 1_{G_i}\right] \le \frac{4}{\mathsf{M}} \frac{\mathrm{Var}(g_i w_t)}{\mathbb{E}_U[g_i w_t]^2} \le \frac{4e^4}{\mathsf{M}} \frac{\mathbb{E}_U[w_t^2]}{\mathbb{E}_U[w_t]^2}. \tag{46}$$

Now, note that $G_i^c = \left\{\left|\frac{\hat{A}_i}{A_i} - 1\right| > \frac{1}{2}\right\}$. This is easily bounded using Chebyshev's inequality:

$$\mathbb{P}(G_i^c) \le \frac{4\mathrm{Var}(A_i)}{A_i^2} \le \frac{4e^4}{\mathsf{M}} \frac{\mathbb{E}_U[w_t^2]}{\mathbb{E}_U[w_t]^2}. \tag{47}$$

On the event G_i^c, we use a universal bound for A_i. Recall the definition of w_t from (43). Note that

$$S_{t-1}(p) = \sum_{s=1}^{t-1} \mathrm{CRPS}(F_s^{(p)}, y_s) \in [0, (b-a)(t-1)]\,,$$

since $0 \le \mathrm{CRPS}(F, y) \le (b-a)$ for any CDF F concentrated on $[a, b]$ and $y \in [a, b]$. Therefore, we have uniformly that

$$1 \ge w_t(p) \ge \exp\left(-2(t-1)\right). \tag{48}$$

As a consequence of the above and the definition (44) of $A_i, \hat{A}_i$ we have

$$\frac{\hat{A}_i}{A_i} \in \left[e^{-2t}, e^{2t}\right] \implies \left|\log \frac{\hat{A}_i}{A_i}\right| \le 2t.$$

Combining the above observation with (47),

$$\mathbb{E}\left[\log^2 \frac{\hat{A}_i}{A_i} 1_{G_i^c}\right] \le \frac{16e^4 t^2}{\mathsf{M}} \frac{\mathbb{E}_U[w_t^2]}{\mathbb{E}_U[w_t]^2}.$$

Finally, combining the above equation with (46),

$$\mathbb{E}\left[\log^2 \frac{\hat{A}_i}{A_i}\right] \le \frac{4e^4(1+4t^2)}{\mathsf{M}} \frac{\mathbb{E}_U[w_t^2]}{\mathbb{E}_U[w_t]^2} \le \frac{4e^4(1+4t^2)}{\mathsf{M}} \frac{1}{e^{-4(t-1)}} = \frac{4e^{4t}(1+4t^2)}{\mathsf{M}},$$

where we used (48) in the final inequality. Inserting the above bound for $i = 0, 1$ into (45) yields

$$\mathbb{E}\left[(\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2\right] \le \frac{4e^{4t}(1+4t^2)}{\mathsf{M}}.$$

Finally, inserting this bound into (41),

$$\mathbb{P}(K > \delta) \le \frac{\sum_{t=1}^T \int_a^b \mathbb{E}\left[(\hat{F}_t(x) - \hat{F}_{\mathsf{M},t}(x))^2\right] dx}{\delta} \le \frac{4(b-a)}{\delta \mathsf{M}} \sum_{t=1}^T e^{4t}(1+4t^2),$$

completing the proof. □

We end this section with a detailed comparison with (Freund, 1996, Theorem 4). The optimal rate obtained by them is $\frac{1}{4}\ln T$, while our worst case bound gives 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 (21). 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 $F_{0,t} = 1_{\{x \le 0\}}$ while $F_{1,t} = 1_{\{x \le 1\}}$. Let $y_i, i \in [T]$ be the observations up till time T. One easily computes from (1) 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 (Freund, 1996), 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 1, 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 (Freund, 1996) and would also be expected to occur in a number of situations, including when p^* is guaranteed to strictly lie inside the simplex.