Fix the double penalty in data-driven forecasting by modifying the loss function

Christopher Subich, Syed Zahid Husain, Leo Separovic, Jing Yang

Christopher.Subich@ec.gc.ca
ECCC (MRD/RPN-Atm)
Dorval, Québec

OceanPredict
Advancing the science of ocean prediction

OceanPredict AI-TT
April 14, 2026
Montreal, Canada

1. Introduction
2. The problem of smoothing
3. Adjusting mean squared error
4. Results
5. Conclusions & extensions

Introduction

- "Deterministic AI models produce smooth forecasts" is conventional wisdom
- Why? Deterministic physics-based forecasts don't inherently smooth predictions.
- Conventional wisdom: it's caused by the Mean Squared Error (MSE) loss
  - Conventional corollary: give up and use an ensemble system
- This talk: why does that happen, and how can we fix it?
  - Conventional wisdom is right: MSE is the problem
  - Conventional corollary is not quite right: we can mostly *fix* MSE
  - More detail in full paper, `https://arxiv.org/abs/2501.19374`
- This presenter: Christopher Subich, generally working on the atmosphere
  - Been working on data-driven forecasting of weather for about three years
  - Prior to that, primarily working on the dynamical core, with academic background in geophysical fluid dynamics & internal wave processes

Mean Squared Error, qualitatively

- MSE suffers from the well-known double penalty effect
  - A model that correctly predicts an eddy but places it at the wrong point is penalized for both a false positive and a false negative
  - Net result is that an MSE-optimized model tries to predict the *conditional mean* of all plausible futures
  - Mean-optimal prediction is *a* choice, but not the only choice
  - We “get away” with MSE-based evaluation for traditional models because they’re tightly constrained towards realism
- Alternatives are less unified:
  - Pooling-based loss functions might capture extremes, but fail to control mean behaviours
  - Probabilistic scoring rules (e.g. CRPS) can induce distribution matching, but these multiply training costs with ensemble forecasts
- Root problem is that data-driven model training requires a single, differentiable, scalar loss function for optimization, but “we want good and realistic forecasts” is fuzzy and multi-faceted

Mean Squared Error, quantitatively

- Tale of two random variables:
  - Analysis: $A = \mathcal{N}(0, \sigma_A^2)$ ("ground truth" and target)
  - Prediction has some standard deviation (σ_P) and is partially correlated (ρ) to the analysis
    - $P = \sigma_P(\rho\sigma_A^{-1}A + \sqrt{1-\rho^2}\mathcal{N}(0,1))$
- Expected $\mathrm{MSE}(P, A) = \mathbb{E}((P-A)^2)$

$$\begin{aligned}\mathrm{MSE}(P, A) &= \mathbb{E}(P^2) + \mathbb{E}(A^2) - 2\,\mathbb{E}(AP) \\ &= \sigma_P^2 + \sigma_A^2 - 2\rho\sigma_A\sigma_P\end{aligned}$$

- $\mathrm{MSE}(P, A) = 0$ iff $\sigma_P\sigma_A^{-1} = \rho = 1$, but the future is inherently unpredictable
- For *fixed* $\rho < 1$, MSE is minimized when $\sigma_P = \rho\sigma_A$: variance reduction
- Mathematical expression of "predict the conditional mean" under finite model capacity

Extension to spectra

- This general result holds for fields (full model outputs) as well as single variables
- It also holds for any linear combination of outputs that satisfies Parseval's theorem
  - Usually seen in Fourier space: $\sum x_i^2 = \sum_k |\alpha_x(k)|^2$ for $\alpha_x(k)$ as the Fourier coefficients of x
  - Also applies to Spherical harmonic decomposition, where $x(i,j) = \sum_k \sum_l \alpha_x(k,l)\, \mathrm{Y}_k^l(i,j)$
  - Privileged position of the atmosphere vs the ocean, where we can pretend we have data over the whole sphere
- For a single model output (p) and analysis (a):

$$\begin{aligned}\mathrm{MSE}(p,a) &= \sum_i \sum_j \mathrm{dA}(i,j)(p(i,j) - a(i,j)^2) \\ &= \sum_k \sum_l \alpha_p^2(k,l) + \alpha_a^2(k,l) - 2\alpha_p(k,l)\alpha_a(k,l)\end{aligned}$$

Extension to spectra – Power spectral density & coherence

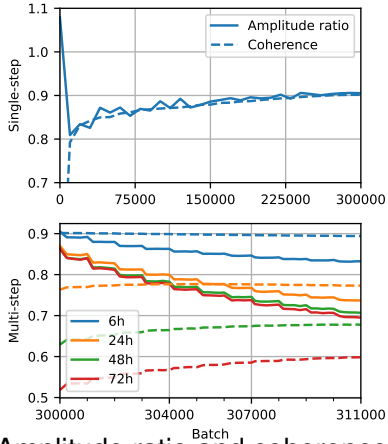
$$\mathrm{MSE}(p,a) = \sum_k \sum_l \alpha_p^2(k,l) + \alpha_a^2(k,l) - 2\alpha_p(k,l)\alpha_a(k,l)$$

- Sum over zonal wavenumbers to get helpful quantities:
  - Power spectral density: $\mathrm{PSD}_k(x) = \sum_l \alpha_x(k,l)^2$, analog of variance
  - Coherence: $\mathrm{Coh}_k(x,y) = \dfrac{\sum_l \alpha_x(k,l)\alpha_y(k,l)}{\sqrt{\mathrm{PSD}_k(x)\,\mathrm{PSD}_k(y)}}$, analog of correlation
- Rewrite MSE:

$$\mathrm{MSE}(p,a) = \sum_k \mathrm{PSD}_k(p) + \mathrm{PSD}_k(a) - 2\sqrt{\mathrm{PSD}_k(p)\,\mathrm{PSD}_k(a)}\,\mathrm{Coh}_k(p,a)$$

- Minimized when the amplitude ratio $\sqrt{\mathrm{PSD}_k(p)\,\mathrm{PSD}_k(a)^{-1}}$ matches the coherence

Smoothing during training



Amplitude ratio and coherence of T850 (wn 100) for 1° training

- This intuition is prescriptive: the amplitude ratio tracks coherence during the whole training process
- Relationship clearest with single-step forecasts, but still directionally true with multi-step forecasts and time-averaged loss function
- Effective resolution is a function of training length, making MSE-based verification *extremely* challenging for AI models
- Smoothing (easy) adjusts more rapidly than coherence/skill (hard)

Adjusting MSE

$$\mathrm{MSE}(p,a) = \sum_k \mathrm{PSD}_k(p) + \mathrm{PSD}_k(a) - 2\sqrt{\mathrm{PSD}_k(p)\,\mathrm{PSD}_k(a)}\,\mathrm{Coh}_k(p,a)$$

- MSE thoroughly mixes spectral amplitudes and coherence (activity and correlation)
- But can be rewritten as an amplitude-only term plus a mixed term
- Optimal smoothing is governed exclusively by this geometric mean term
- . . . so replace it to break that optimum

Adjusting MSE

$$\mathrm{MSE}(p,a) = \sum_k (\sqrt{\mathrm{PSD}_k(p)} - \sqrt{\mathrm{PSD}_k(a)})^2 + 2\sqrt{\mathrm{PSD}_k(p)\,\mathrm{PSD}_k(a)}(1 - \mathrm{Coh}_k(p,a))$$

- MSE thoroughly mixes spectral amplitudes and coherence (activity and correlation)
- But can be rewritten as an amplitude-only term plus a mixed term
- Optimal smoothing is governed exclusively by this geometric mean term
- . . . so replace it to break that optimum

Adjusting MSE

$$\mathrm{MSE}(p,a) = \sum_k (\sqrt{\mathrm{PSD}_k(p)} - \sqrt{\mathrm{PSD}_k(a)})^2 + 2\sqrt{\mathrm{PSD}_k(p)\,\mathrm{PSD}_k(a)}(1 - \mathrm{Coh}_k(p,a))$$

- MSE thoroughly mixes spectral amplitudes and coherence (activity and correlation)
- But can be rewritten as an amplitude-only term plus a mixed term
- Optimal smoothing is governed exclusively by this geometric mean term
- . . . so replace it to break that optimum

Adjusting MSE

$$\mathrm{AMSE}(p,a) = \sum_k (\sqrt{\mathrm{PSD}_k(p)} - \sqrt{\mathrm{PSD}_k(a)})^2 + 2\,\max(\mathrm{PSD}_k(p), \mathrm{PSD}_k(a))(1 - \mathrm{Coh}_k(p,a))$$

- MSE thoroughly mixes spectral amplitudes and coherence (activity and correlation)
- But can be rewritten as an amplitude-only term plus a mixed term
- Optimal smoothing is governed exclusively by this geometric mean term
- . . . so replace it to break that optimum

Adjusted mean squared error – Nice properties

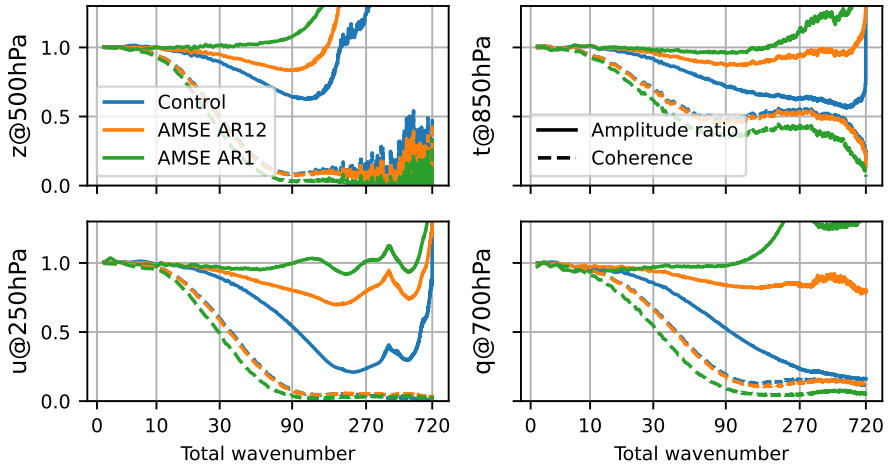
$$\mathrm{AMSE}(p,a) = \sum_k (\sqrt{\mathrm{PSD}_k(p)} - \sqrt{\mathrm{PSD}_k(a)})^2 + 2\max(\mathrm{PSD}_k(p), \mathrm{PSD}_k(a))(1 - \mathrm{Coh}_k(p,a))$$

- Proper scoring rule: $\mathrm{AMSE}(p,a) \geq 0$, with equality iff $p = a$
- Amplitude and coherence components of error share units and potential orders of magnitude
  - No dependence on external/climatological weighting factors
  - Easy to substitute in if available
- Symmetric: $\mathrm{AMSE}(p,a) = \mathrm{AMSE}(a,p)$
- Improving coherence and moving to spectral-amplitude equality independently reduce error

Fine-tuning

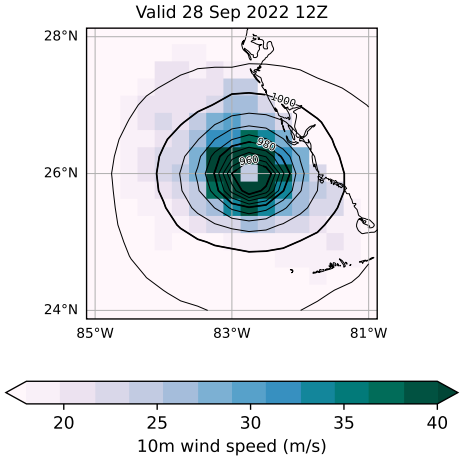
- AMSE was implemented as a loss function in the GraphCast codebase, as a "drop-in" replacement for spatial MSE
  - Spherical harmonic transforms are very fast on GPU, given the lat/lon grid
  - Per-variable and per-level normalizations and error weightings were left unchanged
- Attempt to "fix" GraphCast through fine-tuning:
  - Start with Deepmind ¼°/13L weights, already tuned on HRES initial conditions
  - Train only on HRES data (2016–2021) with a limited GPU budget ($<$ 1 GPU-month) and AMSE loss function
- The fine-tuning process introduces no new data and $<$ 10% additional computation compared to from-scratch training

Effective resolution – Selected variables, +120h



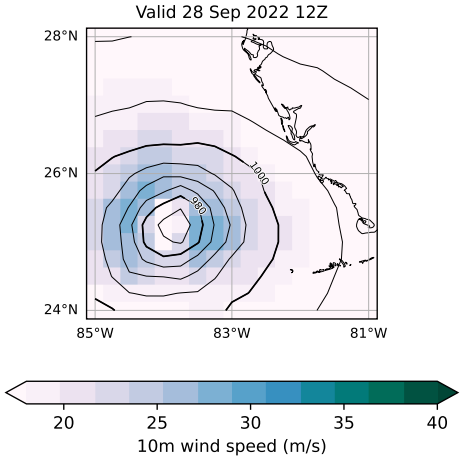
AMSE eliminates dissipation of fine scales, but might get “noisy” at 160km

Tropical Cyclones – Hurricane Ian



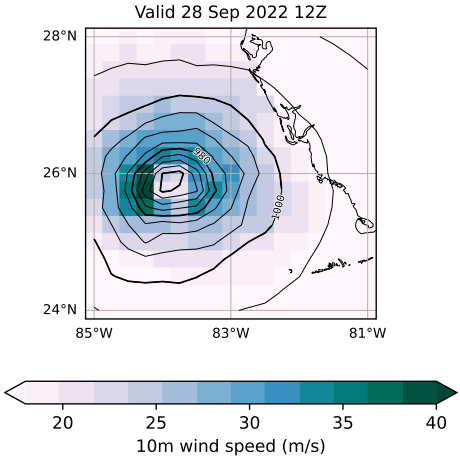
- So this works in theory, but what about in practice?
- Hurricane Ian was the strongest Atlantic tropical cyclone of the 2022 season
- HRES analysis (sea level pressure and surface wind speed) at the cyclone's maximum intensity, just before landfall

Tropical Cyclones – Hurricane Ian – Control



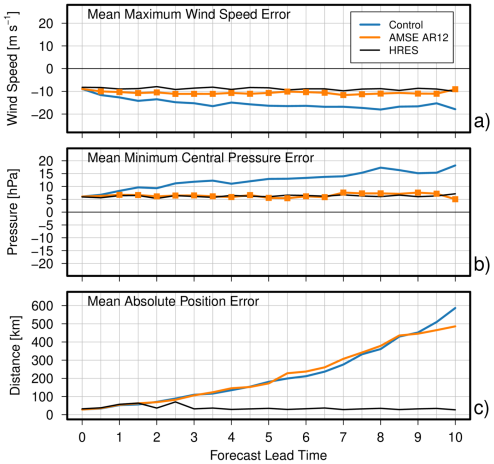
- 5-day forecast (+120h) produced by the control model
- Good prediction of location, but dramatically weakened
- Increased central pressure, reduced surface winds

Tropical Cyclones – Hurricane Ian – AMSE



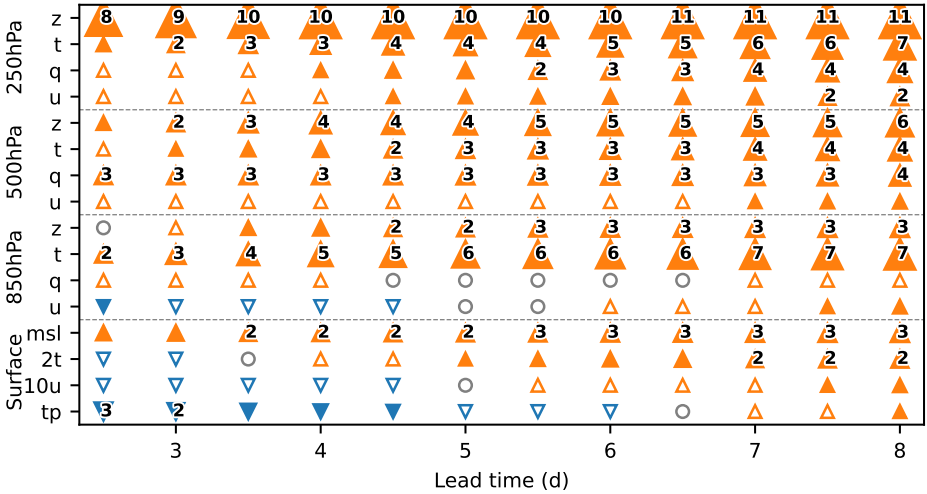
- 5-day forecast (+120h) produced by the fine-tuned model
- Similarly accurate prediction of location
- Much better prediction of intensity

Tropical Cyclones – Systematic evaluation



- Hurricane evaluation against best track database
  - Analysis is imperfect
- June–August 2022
- AMSE training eliminates nearly all biases compared to analysis
- No systematic change in position errors (happy null result)

Lagged ensemble verification – CRPS scorecard



Discussion & limitations

- This loss function gives us the first deterministic ML-NWP models that have anything close to a full spectrum
  - Not necessarily the full story for “physical realism”
  - AMSE loss function assumes that all variability at a particular scale is equivalent
- Advanced loss functions can be implemented as a fine-tuning pass; models learn to smooth or sharpen relatively quickly
- Implications for ocean models:
  - Harder to directly implement because the ocean domain isn’t the full sphere
  - Basic principles don’t really depend on spherical harmonics, however
  - Any multiscale transform like wavelets should suffice, provided there’s a reasonable definition of “equivalence”
  - . . . but fixes based in a loss function are agnostic to model architectures and can apply widely

Future work

- How well does AMSE extend to a true ensemble configuration?
- Swapping out the max-amplitude term for a pre-specified climatological weight makes AMSE a norm on the residual
  - AMSE – a special loss function on the unmodified fields – becomes MSE on the fields after an invertible transform
  - Opens up immediate extensions to more alternative formulations – L1 (MAE-like) loss, CRPS, energy score
  - “Invertible transform” approach should extend to regional multiscale decompositions
- Implications unclear for using AMSE in full training rather than fine-tuning
  - Ideal outcome: model retains realistic variability at all points during training
  - Concern: does smoothing of MSE make learning easier through an automatic “curriculum” of learning large scales first, then fine scales?