Тенденции развития методов усвоения данных наблюдений для прогноза погоды, с акцентом на применении технологий машинного обучения

М.Д. Цырульников

Гидрометцентр России

10 марта 2026 г.

Что такое усвоение данных?

Усвоение данных *даёт оценку состояния системы по данным наблюдений и с использованием математической модели системы*

NB: Оценка состояния включает оценку неопределённости.

- Наблюдения, как правило, имеют недостаточную густоту
- Наблюдаться могут не все интересующие нас параметры системы или не все в равной мере
- Наблюдаться могут не те характеристики системы, которые нас интересуют
- Наблюдения содержат ошибки
- Модель системы тоже неточна

Область применения усвоения данных, помимо метеорологии, включает другие науки о Земле, экологию, космологию, биологию и др.

Специфика метеорологических задач

- Размерность вектора состояния – порядка миллиарда
- Количество наблюдений в сутки – более миллиарда и быстро растёт
- Земная атмосфера относительно хорошо наблюдаема

Наблюдения

Модели наблюдений

1. Прямая модель

$$\mathbf{y} = H(\mathbf{x}) + \varepsilon$$

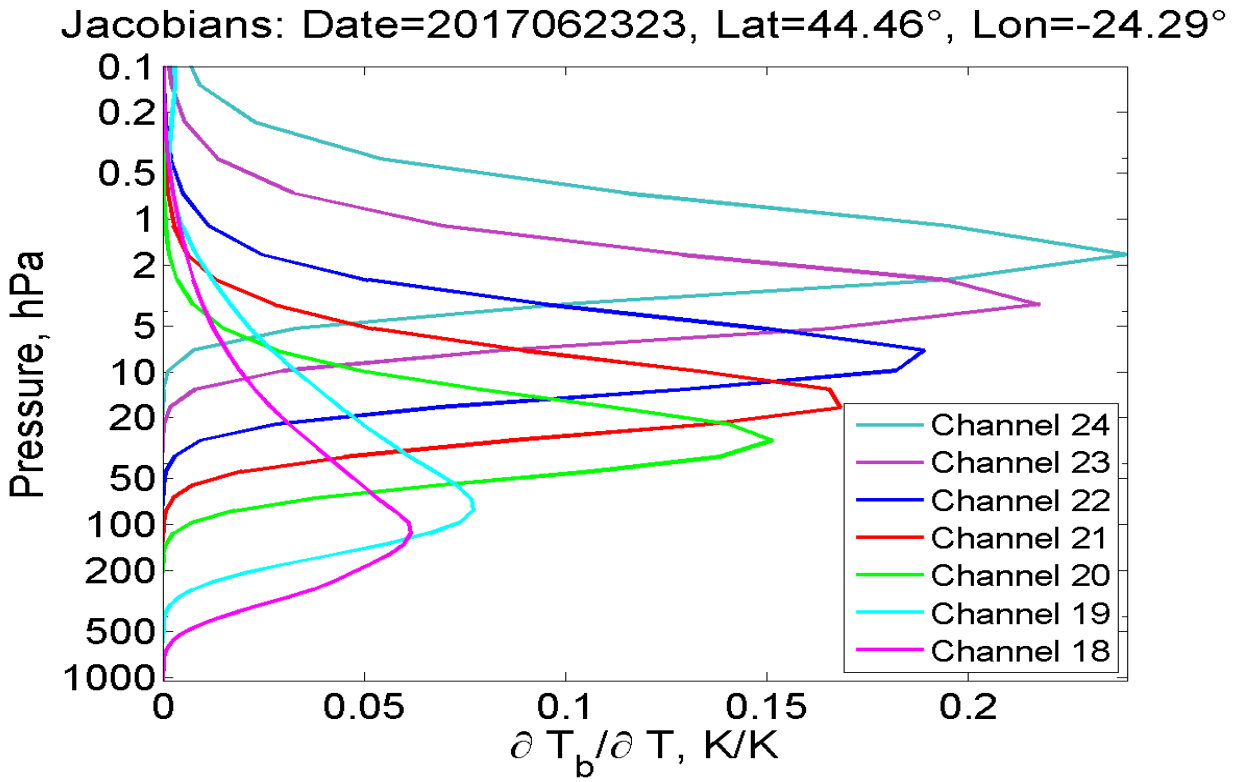
H – модель (оператор) наблюдений,
ε – ошибка наблюдений + ошибка модели $H(.)$

2. Вероятностная модель ошибок $p(\varepsilon \,|\, \mathbf{x})$

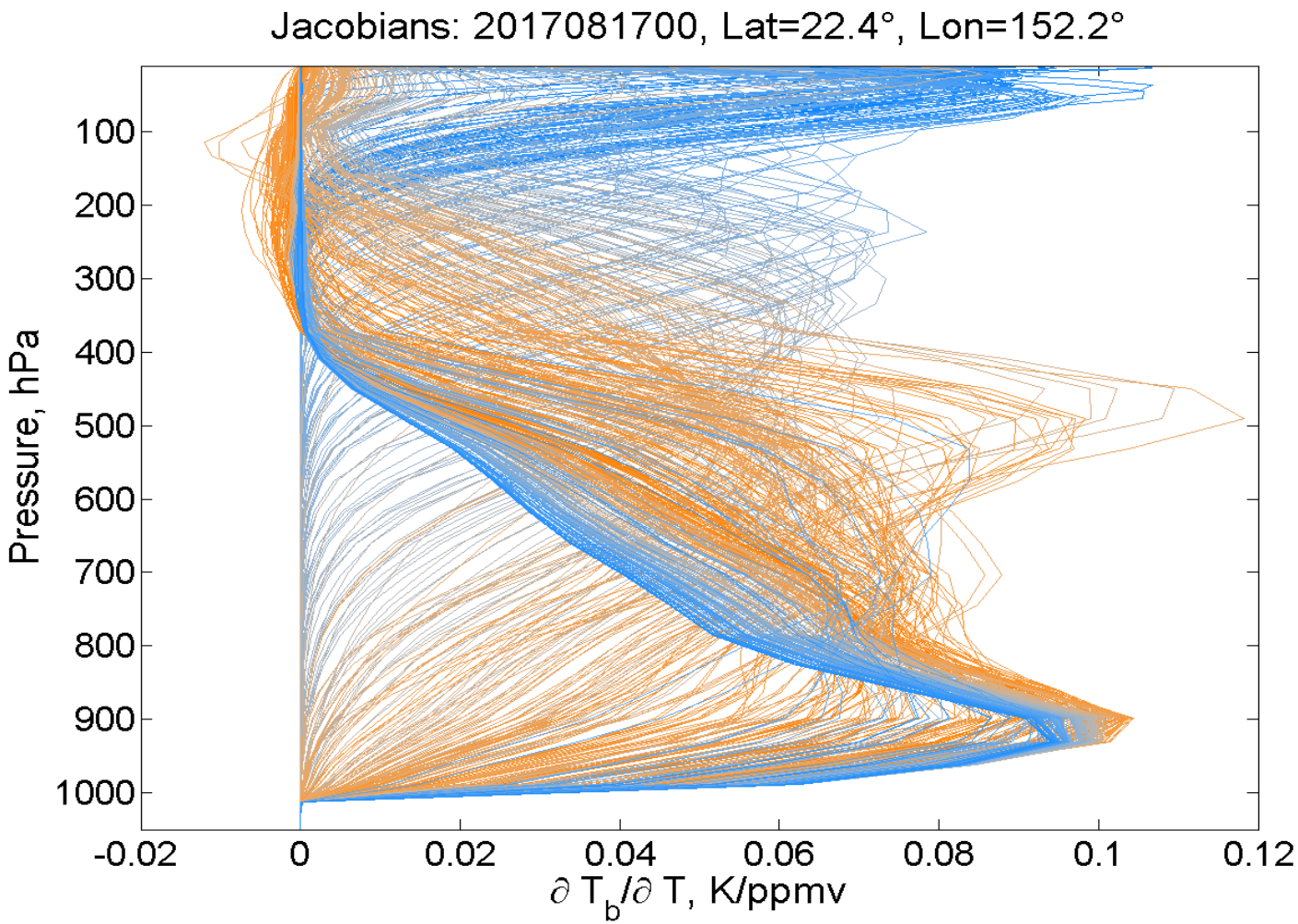
$\Rightarrow$ Функция правдоподобия

$$p(\mathbf{y} \,|\, \mathbf{x})$$

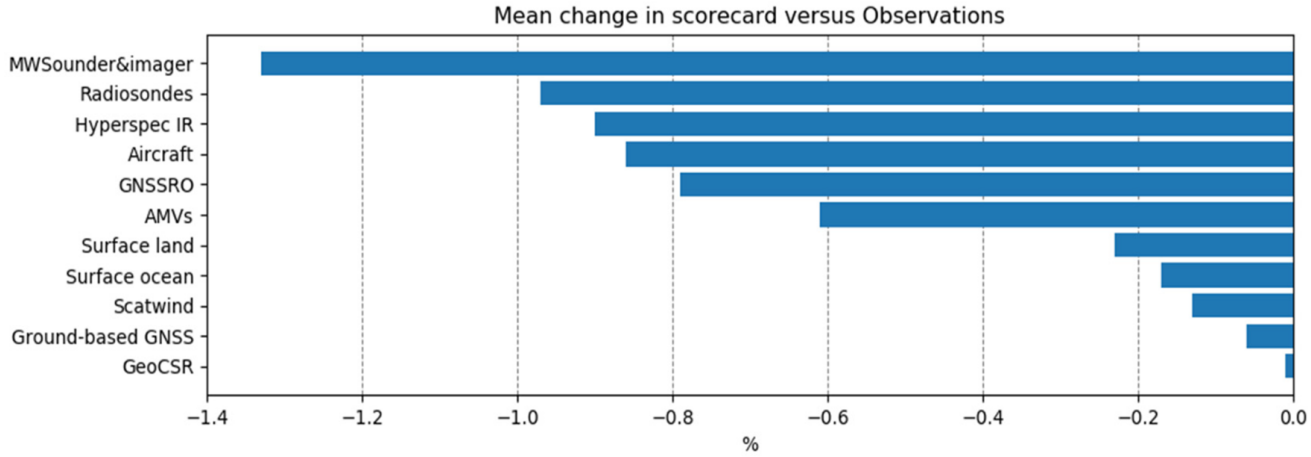
Весовые функции микроволновых наблюдений (МТВЗА-ГЯ)

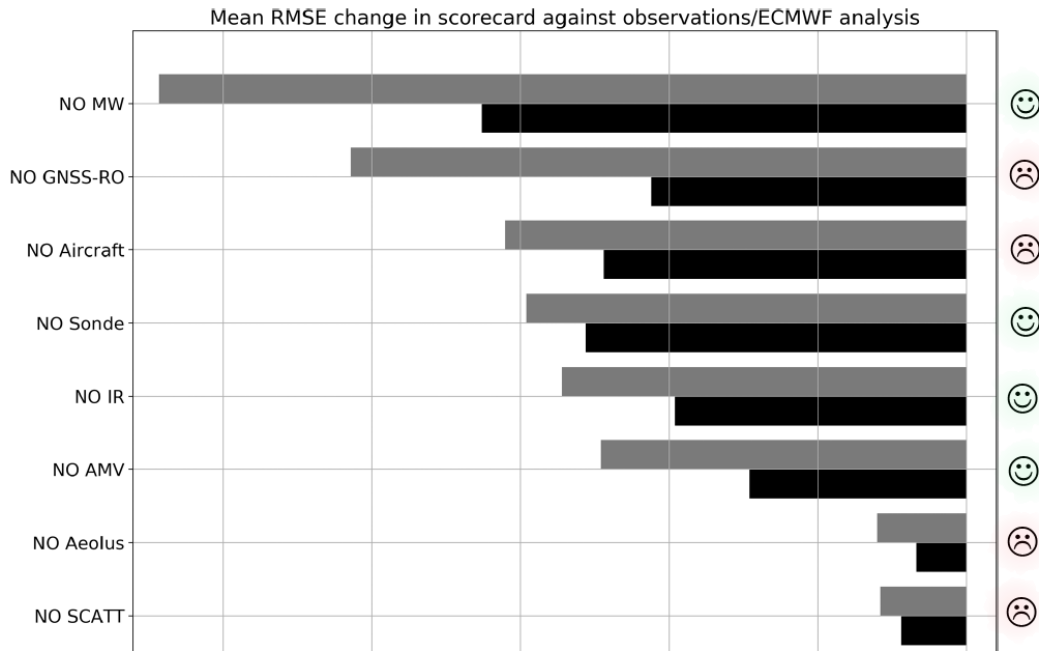


Весовые функции инфракрасных наблюдений (ИКФС-2)



Вклад разных наблюдательных систем в глобальный прогноз, UK Met Office: 2022 г. и 2025 г.





Refs.:
- Assimilation of satellite data in numerical weather prediction. Part II. 2022.
- Observation impact evaluation through data denial experiments in the Met Office... 2025.

Вклад наблюдений: выводы

1. На глобальном масштабе спутниковые наблюдения доминируют, их число растет
2. Контактные наблюдения по-прежнему незаменимы, особенно над сушей и в региональном усвоении
3. В последние годы радикально возрос вклад радиозатменных наблюдений (из-за резкого роста их количества)
4. Вырос вклад самолётных наблюдений (из-за роста их количества)
5. Не “взлетели” crowd-sourced obs

Методы усвоения данных

Классическая парадигма

Байесовский подход. Неопределённость нашего знания истины $\mathbf{x}_k$ моделируем распределением вероятностей. Наша *задача* — оценить это распределение по прошлым и настоящим наблюдениям: $p(\mathbf{x}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_k)$.

У нас есть:

1. Прогностическая модель $\mathcal{F}(\mathbf{x})$ (неточная): $\mathbf{x}_k = \mathcal{F}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k$
2. Модель ошибок прогностической модели: $p(\boldsymbol{\eta}_k \mid \mathbf{x}_{k-1})$

   $\Rightarrow$ Переходная плотность $\boxed{p(\mathbf{x}_k \mid \mathbf{x}_{k-1})}$

3. Наблюдения $\mathbf{y}_k$
4. Оператор наблюдений $\mathcal{H}_k$: $\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\varepsilon}_k$
5. Модель ошибок наблюдений: $p(\boldsymbol{\varepsilon}_k \mid \mathbf{x}_k)$

   $\Rightarrow$ Функция правдоподобия $\boxed{p(\mathbf{y}_k \mid \mathbf{x}_k)}$

В метеорологической реальности модели ошибок весьма неточны.

Циклическое усвоение: теория

1. **Анализ**: от *априорной* плотности $p(\mathbf{x}_k)$ к *апостериорной*:

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) \propto p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1})\, p(\mathbf{y}_k \mid \mathbf{x}_k)$$

2. **Прогноз**: от *апостериорной* плотности $p(\mathbf{x}_k)$ к *прогностической* (априорной на следующем шаге):

$$p(\mathbf{x}_{k+1} \mid \mathbf{y}_{1:k}) = \int p(\mathbf{x}_k \mid \mathbf{y}_{1:k})\, p(\mathbf{x}_{k+1} \mid \mathbf{x}_k)\, \mathrm{d}\mathbf{x}_k$$

...

Результат: на шаге k имеем

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k})$$

Недостаток классической парадигмы: использует ряд упрощающих допущений и предположений.

Циклическое усвоение: практика

1. Распределения вероятностей аппроксимируем многомерным нормальным распределением.
2. Распределение вероятностей характеризуем не плотностью, а выборкой (ансамблем) полей.
3. Оператор наблюдений $\mathcal{H}_k$ $\left[\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \varepsilon_k\right]$ считаем слабо нелинейным и линеаризуем.

Методы оперативного усвоения

1. Вариационные методы: 3D-Var, 4D-Var: $\mathbf{B} = \mathbf{B}_{\text{clim}}$

$$\mathbf{x}^{\text{a}} = \mathbf{x}^{\text{f}} + \mathbf{B}\mathbf{H}^{\mathsf{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathsf{T}} + \mathbf{R})^{-1}(\mathbf{x}^{\text{obs}} - \mathcal{H}(\mathbf{x}^{\text{f}}))$$

2. Ансамблевые методы: EnKF: $\mathbf{B}_{\text{ens}} = (1+\epsilon)\,\mathbf{S} \odot \mathbf{C}$

3. Ансамблево-вариационные методы: EnVar, 4D-Var с ансамблевой статистикой ошибок первого приближения

$$\mathbf{B} = w\mathbf{B}_{\text{ens}} + (1-w)\mathbf{B}_{\text{clim}}$$

Не вполне обоснованные процедуры:

– *оценка* $\mathbf{B}_{\text{clim}}$
– *задание* $\mathbf{B}_{\text{ens}}$
– *смесь* $\mathbf{B}_{\text{clim}}$ *и* $\mathbf{B}_{\text{ens}}$.

– *Вариационные методы: минимизация функционала ошибки* $\Leftrightarrow$ *оценка макс. апостериорной плотности (MAP). Почему это хорошо для нелинейных наблюдений?*

Оперативная практика: дополнительные научно не обоснованные процедуры

= коррекция смещений наблюдений по отношению к прогнозу (который тоже смещён)

= контроль качества наблюдений – повторное использование прогноза

= вертикальная интерполяция/экстраполяция прогноза на наблюдения

Методы машинного обучения в усвоении данных

Нейросети. Цели

- Основываться на данных – там где математическая модель неточна или отсутствует
- Радикально ускорить вычисления (на порядки)
- Тренировать модель по конечному результату

Нейросети. Способы применения в усвоении данных

- Заменить дорогую физическую модель на дешёвую нейросетевую на шаге прогноза
- Заменять блоки анализа: контроль качества, коррекция смещений и т.п. + оператор наблюдений
- Заменить вероятностную парадигму усвоения данных на эмпирический минимизатор ошибок анализа
- Гибридные подходы
- Отказаться вообще от усвоения данных:
  исключить шаг формирования оценки текущего состояния системы и давать (нейронный) прогноз прямо с наблюдений

- (Усвоение данных для обучения нейросетей)

•Нейросетевые прогностические модели

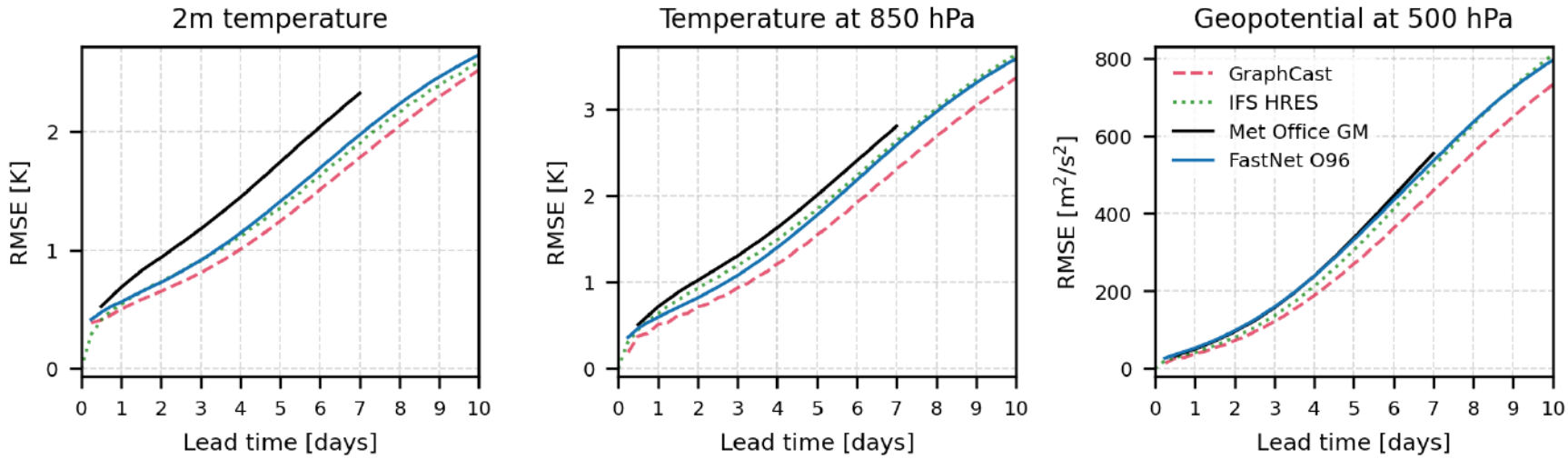


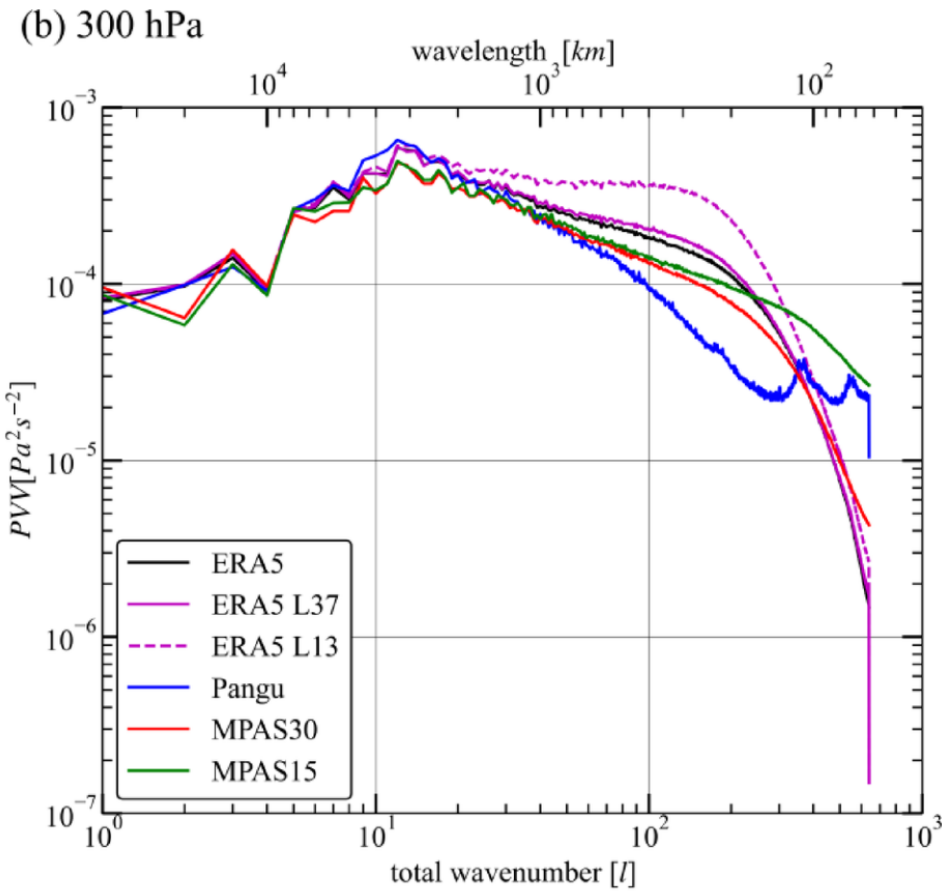
Figure 7: RMSE values as a function of lead time for the Met Office Global Model (black) and fine-tuned FastNet O96 model (blue). IFS-HRES and and GraphCast are also shown for comparison. All models are evaluated in 2022, regridded to 1.5°, and compared to the appropriate ground truth dataset.

Ref.: Daub et al. (UK Met Office + Alan Turing Institute)
Technical overview and architecture of the FastNet Machine Learning weather prediction model. 2025

•(1) Нейросетевая прогностическая модель на шаге прогноза

Нейросетевые модели хорошо предсказывают крупные масштабы (несущие основную энергию) и плохо (пока) предсказывают мезо и мелкие масштабы.

Спектр вертикальной скорости:

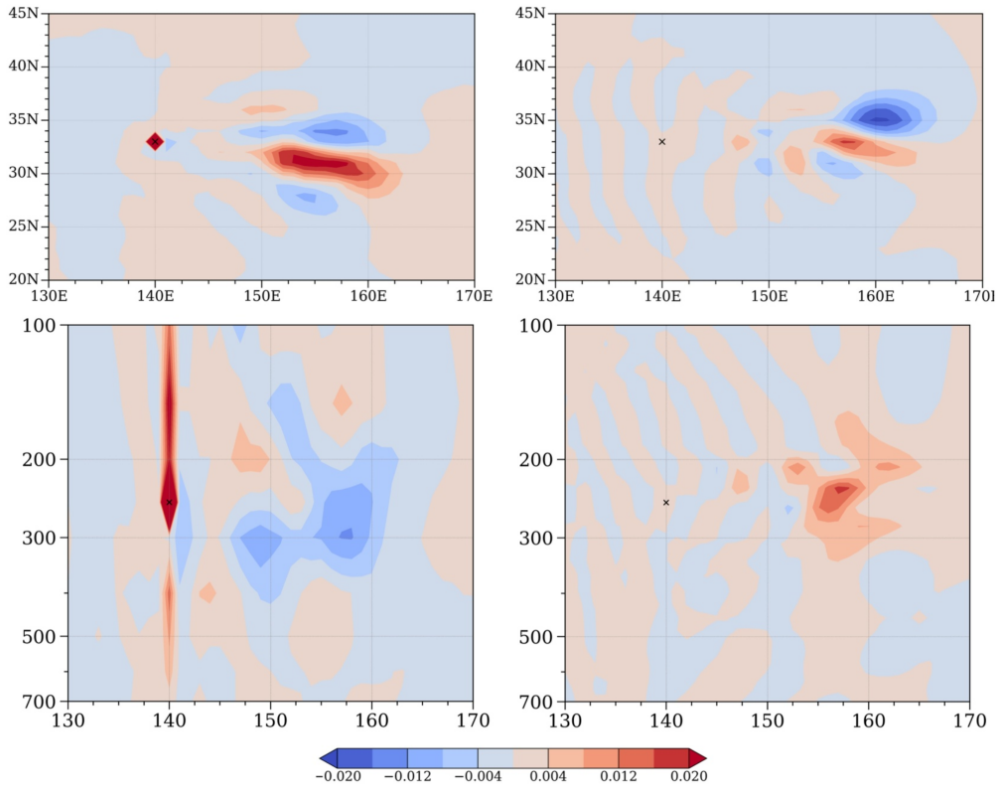


Ref.: Li et al. (China)
Exploring the differences in atmospheric mesoscale kinetic energy spectra between AI based and physics based models. 2025.

Нейросетевая прогностическая модель на шаге прогноза

Нейросетевые модели генерируют ложные возмущения через 6ч линеариз. прогноза:

GraphCast Физ.модель

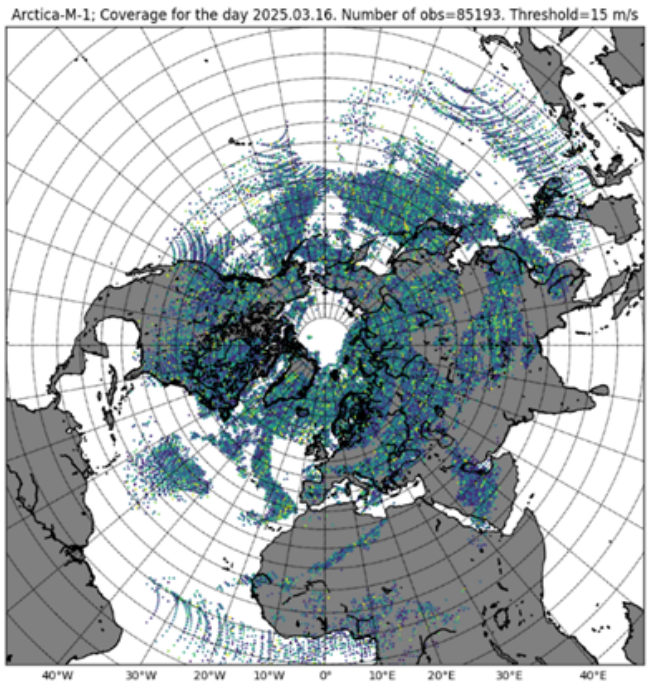


Ложные возмущения ⇒ ложные ковариации ошибок прогноза ⇒ плохой анализ

Ref.: Tian et al. (NOAA)
Evaluating machine learning weather models for data assimilation: Fundamental limitations in tangent linear and adjoint properties. 2026

•(2) Нейросетевые блоки в анализе:
Усвоение спутниковых данных о ветре (AMV).
Нейросетевой оператор наблюдений
(Гайфулин, Цырульников)

Усваиваются результаты восстановления векторов ветра (НИЦ Планета, Хабаровск) по измерениям инфракрасного сенсора МСУ-ГС со спутников Арктика-М на высокоэллиптических орбитах. Покрытие:



Усвоение спутниковых данных о ветре: подход

Наблюдение ветра рассматривается не как “ветер в точке”, а как нелинейный функционал от истинного поля в некотором объёме / по вертикали:

$$u^{\mathrm{obs}}(\theta,\phi) = \mathcal{H}\left(u^{\mathrm{true}}(\cdot,\cdot,\cdot), T_{\mathrm{b}}^{\mathrm{obs}}\right) + \epsilon$$

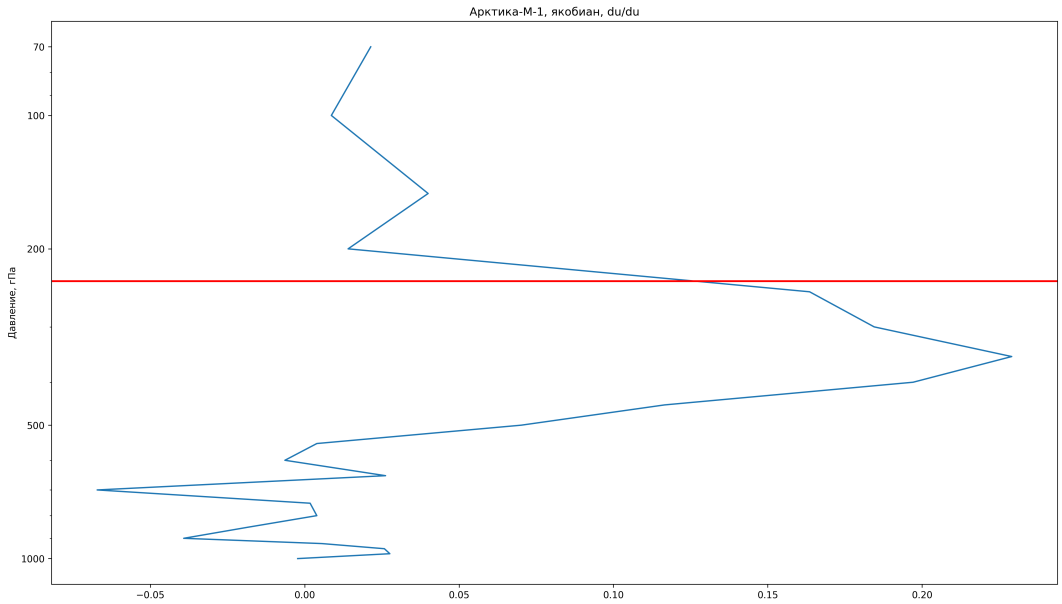
Прямая модель $\mathcal{H}$ – обучаемая.

- Нелокальность оператора отражает природу этих наблюдений
- Коррекция смещений включена в оператор наблюдений
- Независимость от внешнего первого приближения - сейчас используется НСЕП для вертикальной привязки

Реализация: Нейросетевая модель наблюдений, обученная на реальных данных.

Усвоение спутниковых данных о ветре

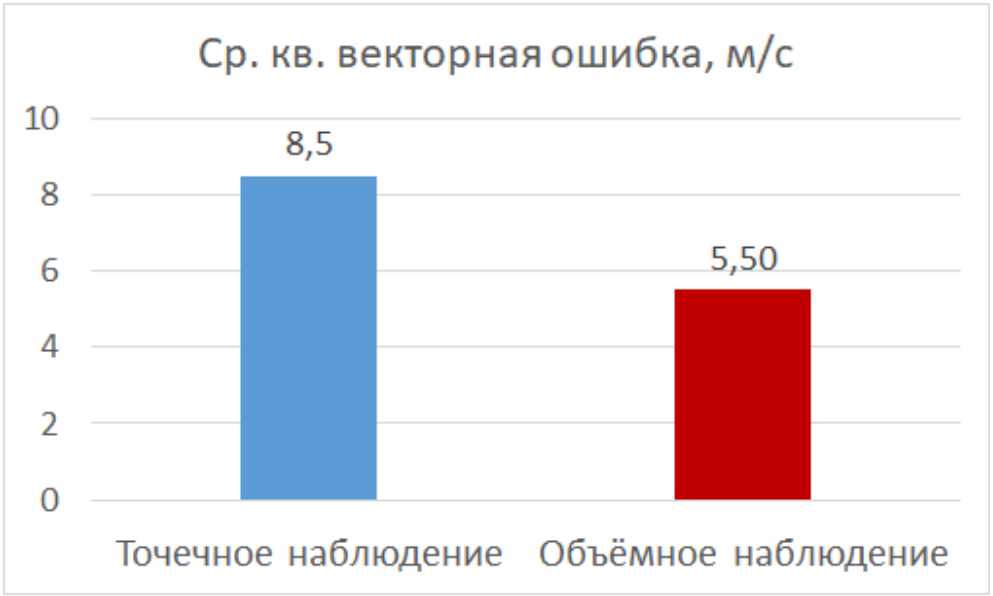
Весовая функция (“якобиан”) $\frac{\partial U^{\mathrm{obs}}}{\partial U_k}$, k – вертикальный уровень



Красная прямая – вертикальная привязка “точечного наблюдения”

Усвоение спутниковых данных о ветре

Среднеквадратичные ошибки точечных и объёмных наблюдений



Для сравнения:

Ср. кв. ошибки векторного ветра AMV (CIMSS, USA) (“точечное наблюдение”): 4.5 м/с (но с использованием прогноза для контроля качества)

Самолётные и радиозондовые наблюдения ветра: 2.5 м/с

(3) Замена вероятностной парадигмы на шаге анализа

1. От традиционного оператора анализа:

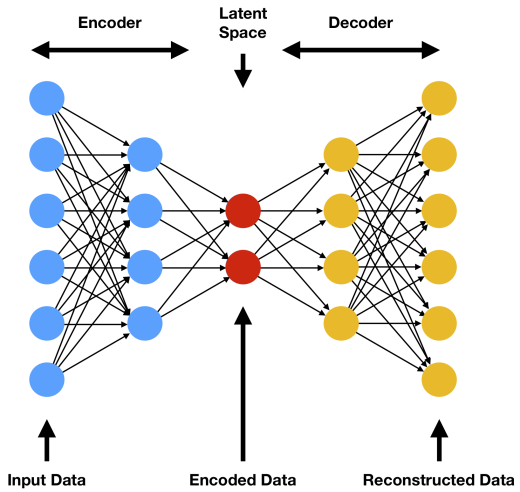
$$\mathbf{x}^{\mathrm{a}}_{\mathrm{traditional}} = \mathbf{x}^{\mathrm{f}} + \mathbf{K} \cdot \left(\mathbf{x}^{\mathrm{obs}} - \mathcal{H}(\mathbf{x}^{\mathrm{f}}) \right)$$

— к обучаемому эмпирическому оператору анализа:

$$\mathbf{x}^{\mathrm{a}}_{\mathrm{neural}} = \mathbf{A}(\mathbf{x}^{\mathrm{f}}; \mathbf{x}^{\mathrm{obs}}, \mathbf{m})$$

2. Diffusion models (ensemble score filters)

3. Latent-space assimilation

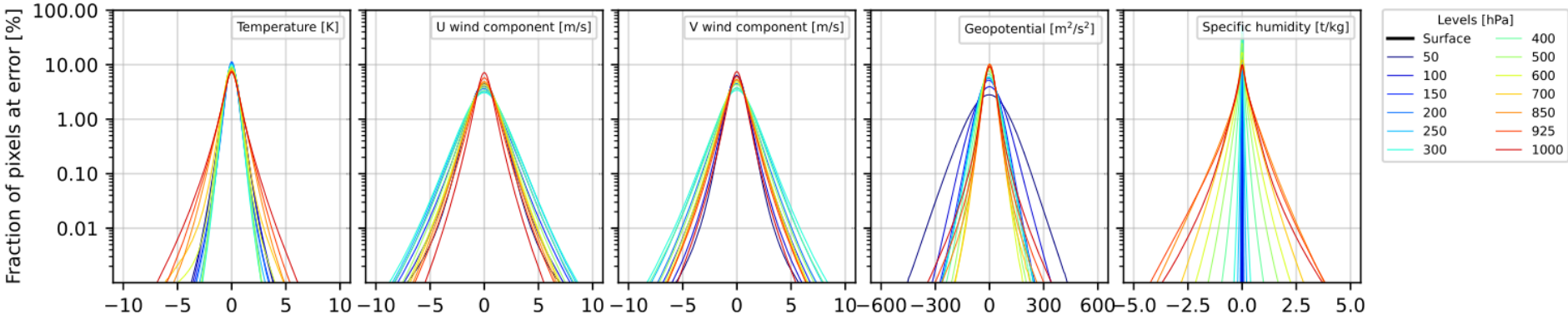


Ошибки восстановления состояния атмосферы из latent space

Appa: Разрешение 0.25 град., 13 уровней.
Архитектура: Автоенкодер + score-based (diffusion model) analysis and forecast.

72ч прогнозы H500: RMSE = 46 m (для сравнения IFS 14 m, AIFS 12 m)

Гистограммы ошибок восстановления:



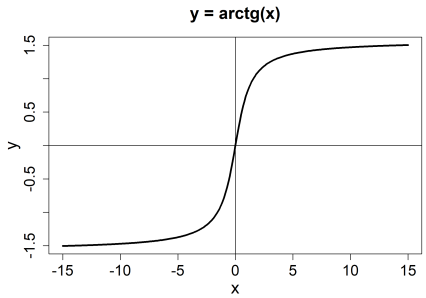
Автоенкодер сжимает в 450 раз (только), при этом
ошибки восстановления соизмеримы как с ошибками прогноза так и ошибками наблюдения

Ref.:
Andry et al. (U Liège)
Appa: Bending Weather Dynamics with Latent Diffusion Models for Global Data Assimilation,2025

Diffusion-model based filters

2D квазигеостр. модель

Сильно нелинейный оператор наблюдений:



Ср. кв. ошибки анализа. Дисперсия ошибок наблюдений =1

Table 1. Time-averaged analysis RMSEs from LETKF and EnSF experiments assimilating 1024 observations (25% of all state variables). Each row represents a different percentage of nonlinear (arctangent) observations in the system.

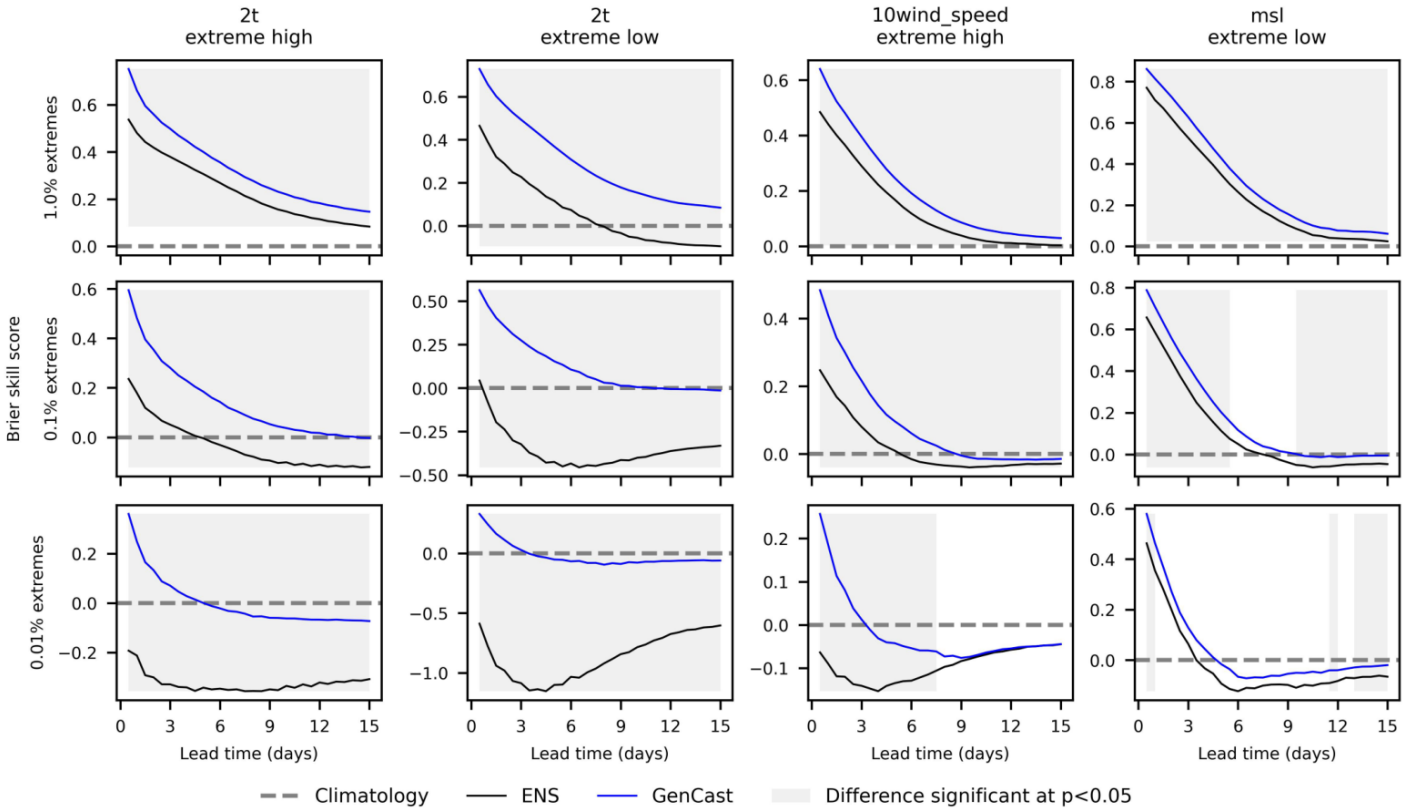
	LETKF			EnSF		
Arctan%	FIXED	FIXED_EVEN	RANDOM	FIXED	FIXED_EVEN	RANDOM
0%	0.716	0.642	0.645	2.756	2.470	2.425
20%	10.117	7.957	9.015	2.780	2.482	2.446
40%	9.885	8.967	9.086	2.792	2.516	2.445
60%	9.632	9.531	9.695	2.797	2.508	2.458
80%	9.440	9.083	10.128	2.798	2.515	2.471
100%	9.209	9.044	9.175	2.836	2.551	2.501

При линейных наблюдениях традиционный ансамблевый фильтр намного точнее.
Нейронный фильтр опережает при очень нелинейных наблюдениях.

Ref.: Xiong et al. (Florida State U)
On the sensitivity of different ensemble filters to the type of assimilated observation networks. 2025

•Ансамбли

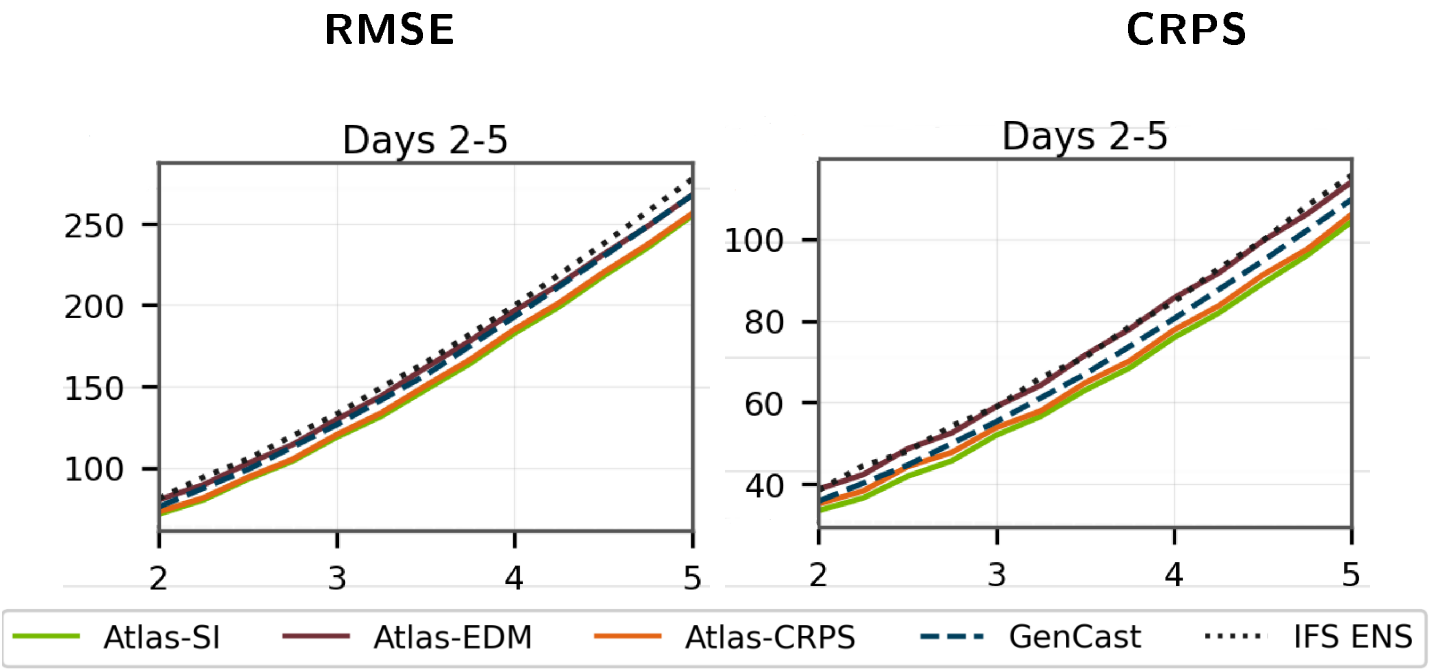
GenCast (стартует с анализа):
Conditional diffusion model. Trained to generate samples from $p(\mathbf{x}_{k+1} \mid \mathbf{x}_k, \mathbf{x}_{k-1})$
Extreme-event verification, Brier skill score (чем больше, тем лучше) – по сравнению с ансамблевым прогнозом ECMWF



Ref.: Price et al. (Google DeepMind)
Probabilistic weather forecasting with machine learning. Nature, 2025.

Ещё ансамбли

ATLAS (стартует с анализа):
Diffusion model. Trained to generate samples from $p(\mathbf{x}_{k+1} \mid \mathbf{x}_k)$
H500 по сравнению с ансамблевым прогнозом ECMWF и GenCast



Ref.: Kossaifi et al. (NVIDIA)
Demystifying Data-Driven Probabilistic Medium-Range Weather Forecasting. 2026.

(4) Гибридные подходы

1. 4D-Var + GAN (Loss function and initial gradient from 4D-Var, Solver using GAN)
2. EnKF (ковариации) + LSTM (прошлое) + MLP
3. EnKF in neural space (updates NN weights and biases)
4. EnKF (lowRes) + CNN (hiRes)
5. EnKF + Reinforcement learning
6. Пред-усвоение наблюдений очень высокого разрешения (CNN)
7. Наш локально стационарный фильтр (GPR + ensembles + NN)

Наша идея –
использовать нейросеть для оптимизации извлечения информации из ансамбля прогнозов –
вместо локализации, инфляции, линейной комбинации с климатическими ковариациями ...

Разработка локально стационарного фильтра (Шен, Цырульников)

•Нестационарная модель поля ошибок прогноза:

$$\xi(x) = \int u(x, \rho(x, y))\, \alpha(y)\, \mathrm{d}y$$

Спектральное представление

$$\xi(x) = \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \sigma_\ell(x)\, \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x)$$

•Оцениватель локальных символов $\sigma_\ell(x)$ по ансамблю с помощью нейросети:

$$\widehat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}_\psi(\mathbf{e})$$

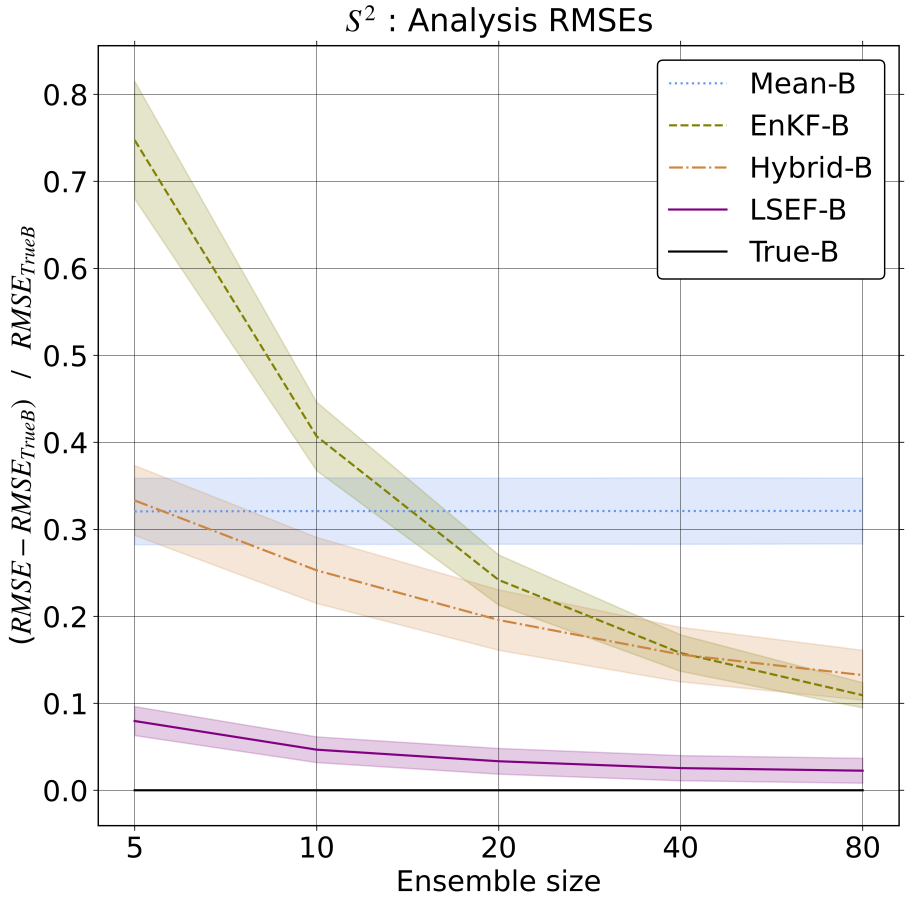
$\mathbf{e}$ – ансамблевая статистика, поточечно сжатая пакетом пространственных фильтров, $\boldsymbol{\psi}$ – параметры сети

Байесовский оцениватель: нейросетевое приближение

1. Априорное распределение искомого параметра $\boldsymbol{\sigma}$
2. Правдоподобие ансамбля $\mathbf{e}$
3. Оцениватель $\boldsymbol{\sigma}_\psi(\mathbf{e})$
4. Функция потерь
$$\mathcal{L}(\boldsymbol{\sigma}_\psi(\mathbf{e}), \boldsymbol{\sigma})$$
5. Байесовский риск:
$$R(\psi) = \mathbb{E}_{\boldsymbol{\sigma},\mathbf{e}}\, \mathcal{L}(\boldsymbol{\sigma}_\psi(\mathbf{e}), \boldsymbol{\sigma})$$
6. Приближение Монте-Карло:
$$\hat{R}(\psi) = \frac{1}{n}\sum \mathcal{L}(\boldsymbol{\sigma}_\psi(\mathbf{e}_{ij}), \boldsymbol{\sigma}_i) \to \min_\psi$$
7. $\Rightarrow$ оценка параметров сети к оптимальной Байесовской оценке.

M.Tsyrulnikov and A.Sotskiy. Regularization of the ensemble Kalman filter using a non-parametric, non-stationary spatial model. Spatial Statistics, 2024, v. 64, N10.

Ср.кв. ошибки анализа (сфера, статический анализ)

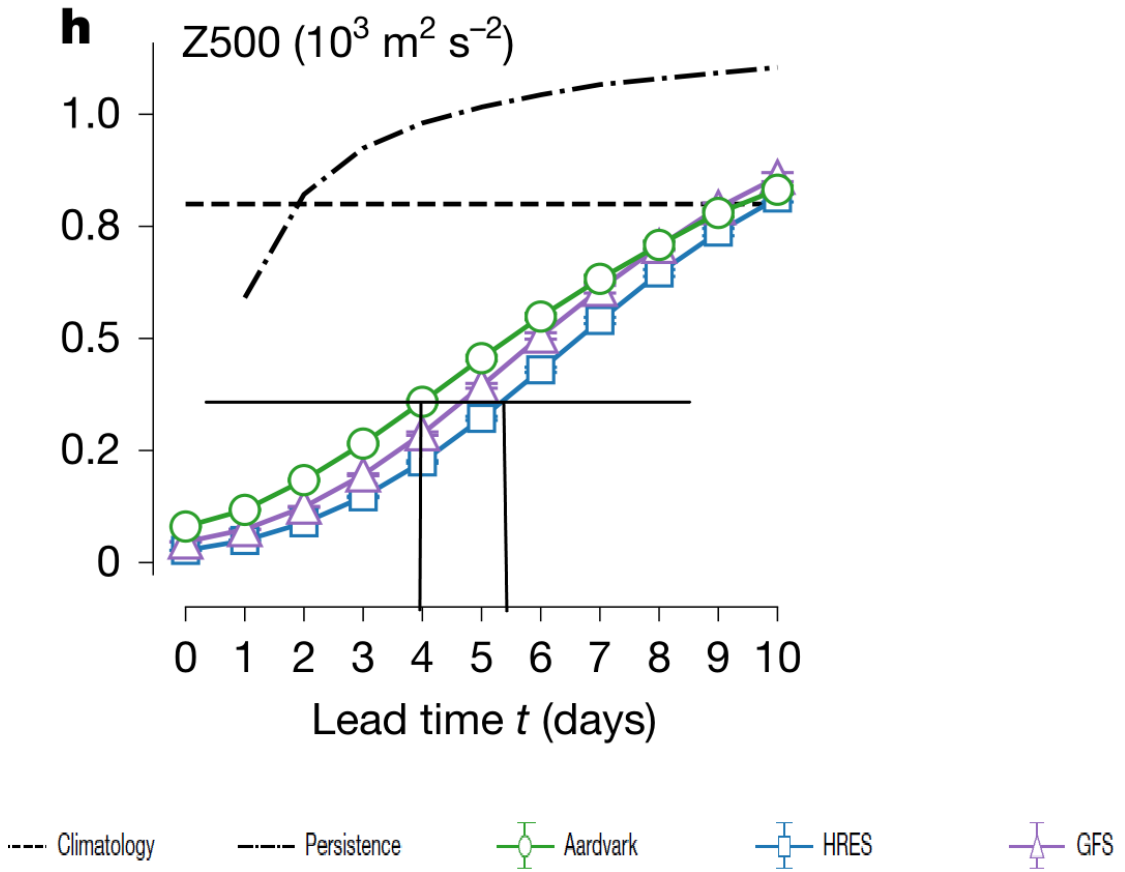


•(5) Полная замена усвоения и прогноза на нейросеть

Нейросеть делает прогноз **непосредственно с наблюдений** + возможно, обучение тоже по наблюдениям.

Aardvark: no assimilation

Obs: in-situ (Rsonde, Synop, Ship) and satellite (MW, IR, Scatt, GEO images).
Obs window: 24h, no cycling. Обучение на реанализе.
Verification on ERA-5.

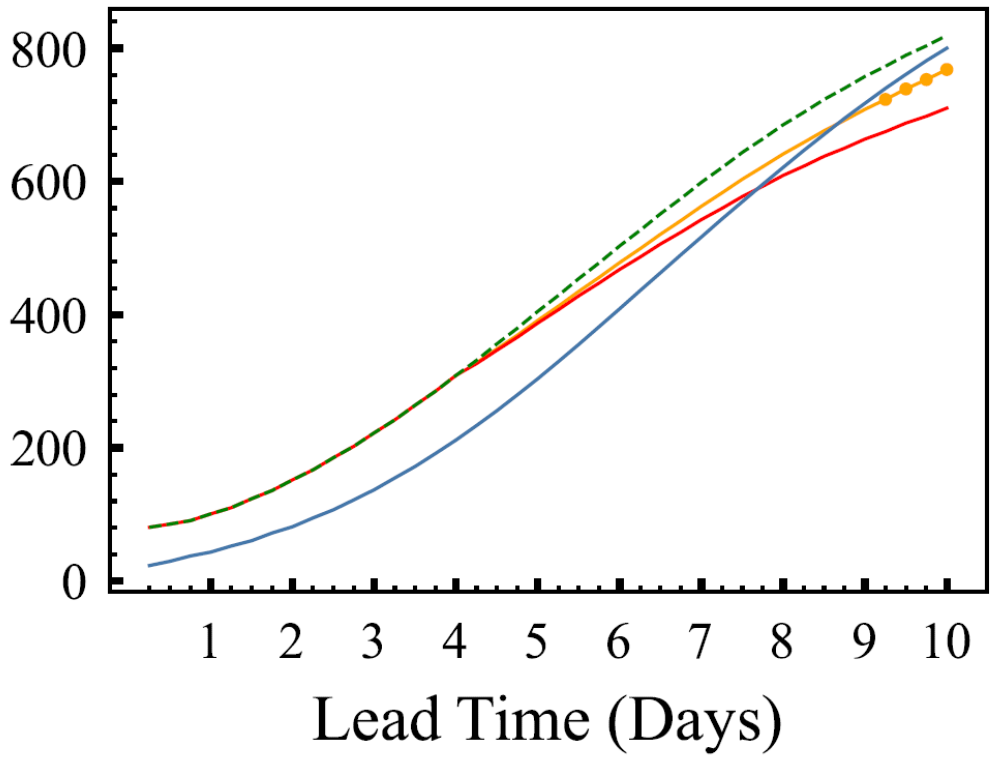


Ref.: Allen et al. (U Cambridge)
End-to-end data-driven weather prediction. Nature, 2025

FuXi Weather: a Data-to-Forecast model (Китай): no assimilation

Обучение на реанализе.
6-h cycle

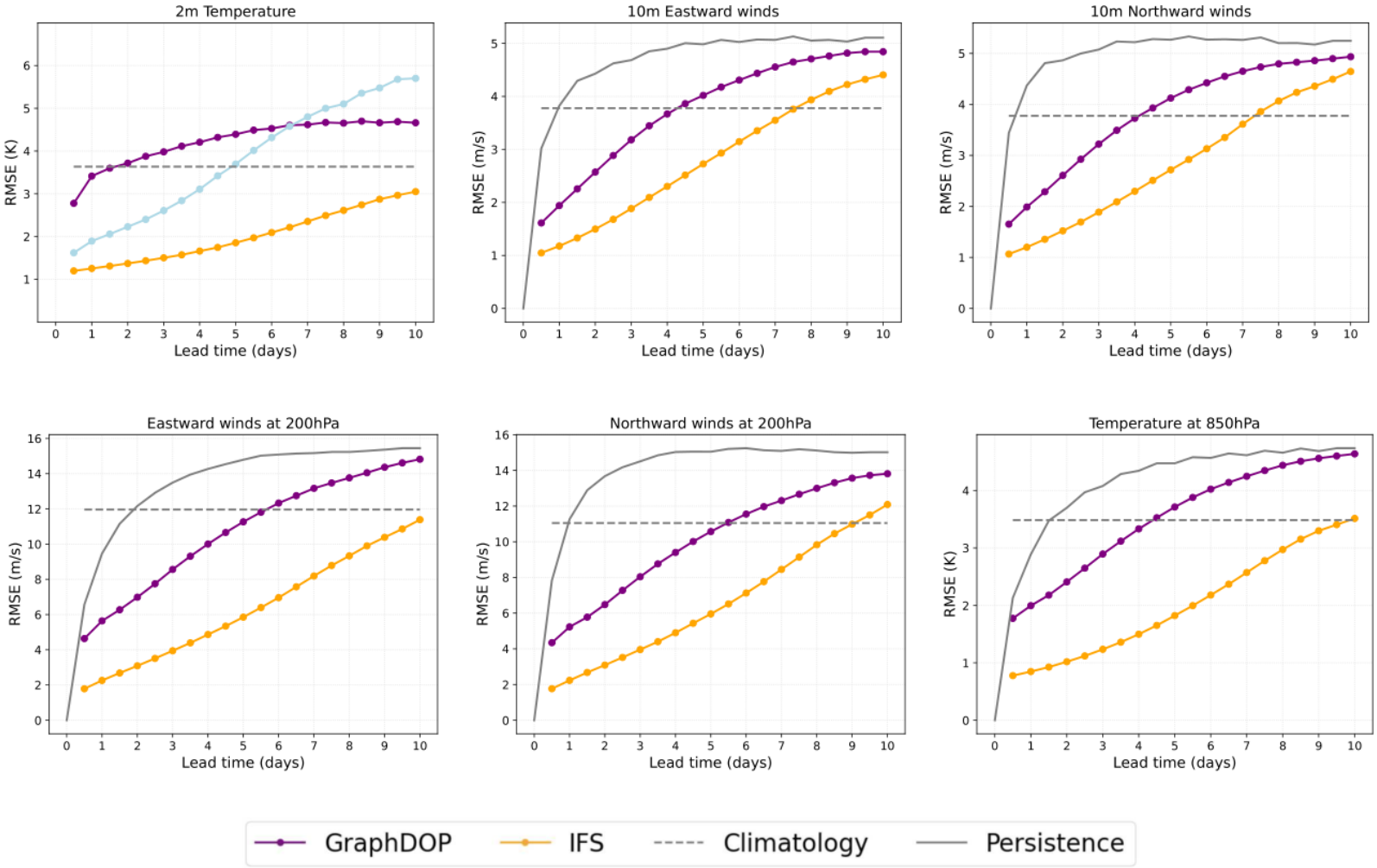
RMSE H500
Синяя линия - ECMWF
Остальные – версии FuXi-Weather



Ref.: Sun et al. (China)
A data-to-forecast machine learning system for global weather. Nature Communications, 2025.

Obs2obs. GraphDOP (ECMWF): no assimilation

Обучение без реанализа, прямо по наблюдениям. No cycling

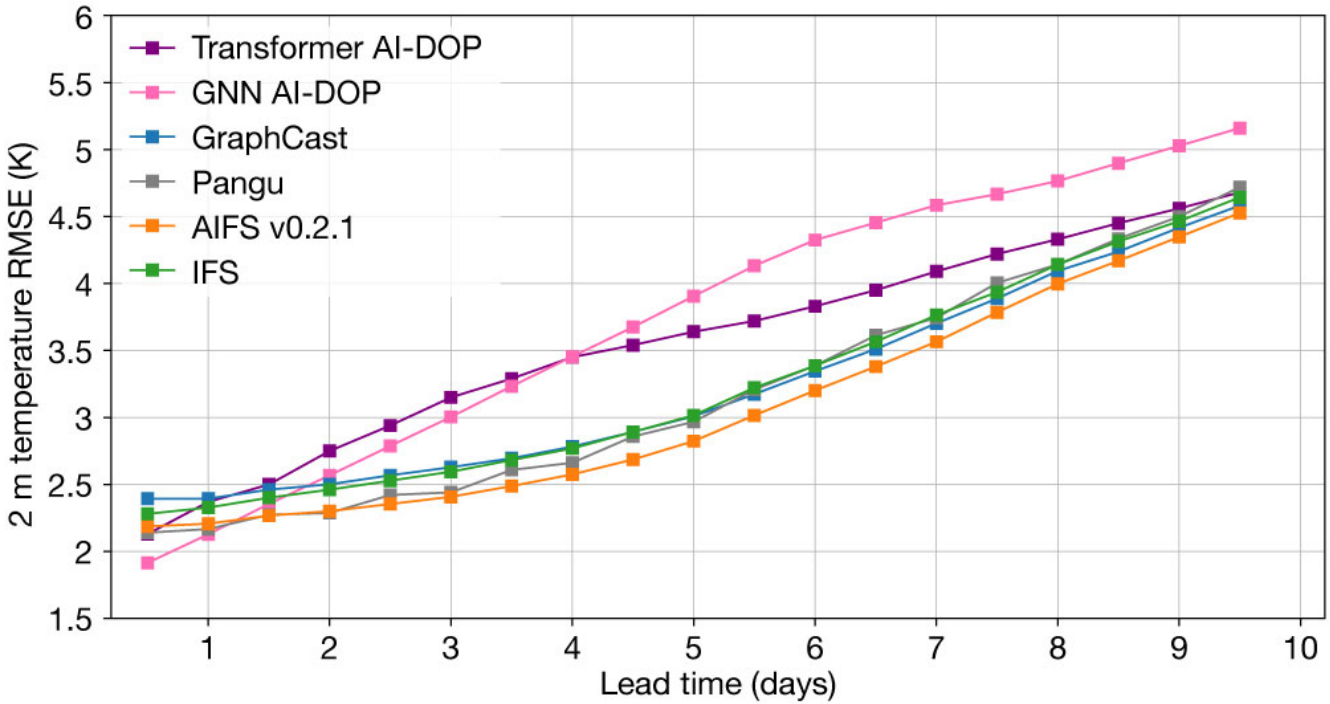


Ref.: Alexe et al. (ECMWF)

GraphDOP: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations, 2024

GraphDOP ⇒ AI-DOP (ECMWF) 2025: no assimilation

12-h observation window. No cycling



Выводы по попыткам полной замены усвоения и прогноза на нейросеть

- Если обучение – по реанализу, то слишком гладкие поля
- Если обучение – по наблюдениям, то принципиально возможен прогноз не всех полей
- Пока нет ансамблей по наблюдениям
- Очень дорогое обучение
- Требуется огромное число параметров сети
- По точности прогнозов пока значительно уступает традиционным методам

= Преимущества нейросетевых подходов

- Основываются на данных, а не на упрощённых теоретических моделях
- Вычислительная эффективность
- Нелинейность
- Универсальность: от текста, речи, картинок, видео до прогноза погоды — однотипные технологии
- Оптимизиция производится непосредственно по конечному результату

= Недостатки нейросетевых подходов

- Затратное обучение
- Необходимость постоянного переобучения при изменениях в системе :
  -новые виды наблюдений
  -изменения в составе наблюдений
  -дрейф параметров спутниковых радиометров и т.п.
- Неустойчивость результатов при выходе “за пределы” обучающей выборки
  -Adversarial attacks
  -Резкие изменения климатической системы (супер-вулкан)
- “Чёрный ящик”. Неинтерпретируемость
- Эмпирический характер построения и настройки сети

Выводы. 1. Наблюдения

Прогресс в оперативных атмосферных наблюдениях релевантных для прогноза погоды происходит в следующих направлениях:

- Постепенный рост количества и качества радиационных спутниковых измерений
- Существенный рост количества самолётных наблюдений
- Взрывной рост количества радиозатменных измерений (микро-спутники)

Выводы. 2. Методы усвоения

Прогресс в методах усвоения данных для прогноза погоды происходит преимущественно путем привлечения методов машинного обучения:

- Пред-обработка наблюдений
- Модели (операторы) наблюдений
- Нейросетевая модель на шаге прогноза в цикле усвоения данных
- Замена оператора анализа, основанного на вероятностный моделях неопределённости прогноза и наблюдений, на обучаемый эмпирический оператор
- Гибридные методы
- Нейросетевой прогноз прямо с наблюдений

Наш подход – основываться на вероятностных моделях неопределенности и привлекать нейросети:

- **в тех компонентах схемы усвоения данных, которые наименее обоснованы/обосновываемы**
- **для ускорения вычислений.**

ВОПРОСЫ, ПОЖАЛУЙСТА!

Оперативная верификация ECMWF https://charts.ecmwf.int/

Scores of forecasts of upper-air parameters by experimental machine learning models