Hybrid metaheuristic optimization for variational data assimilation in turbulence reanalysis

Grzegorz Zakrzewski[1,2], Jacek Mańdziuk[2,3]

[1]Institute of Meteorology and Water Management – National Research Institute, Warsaw, Poland
[2]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
[3]Faculty of Computer Science, AGH University of Krakow, Krakow, Poland

Sharing is encouraged

This presentation participates in OSPP

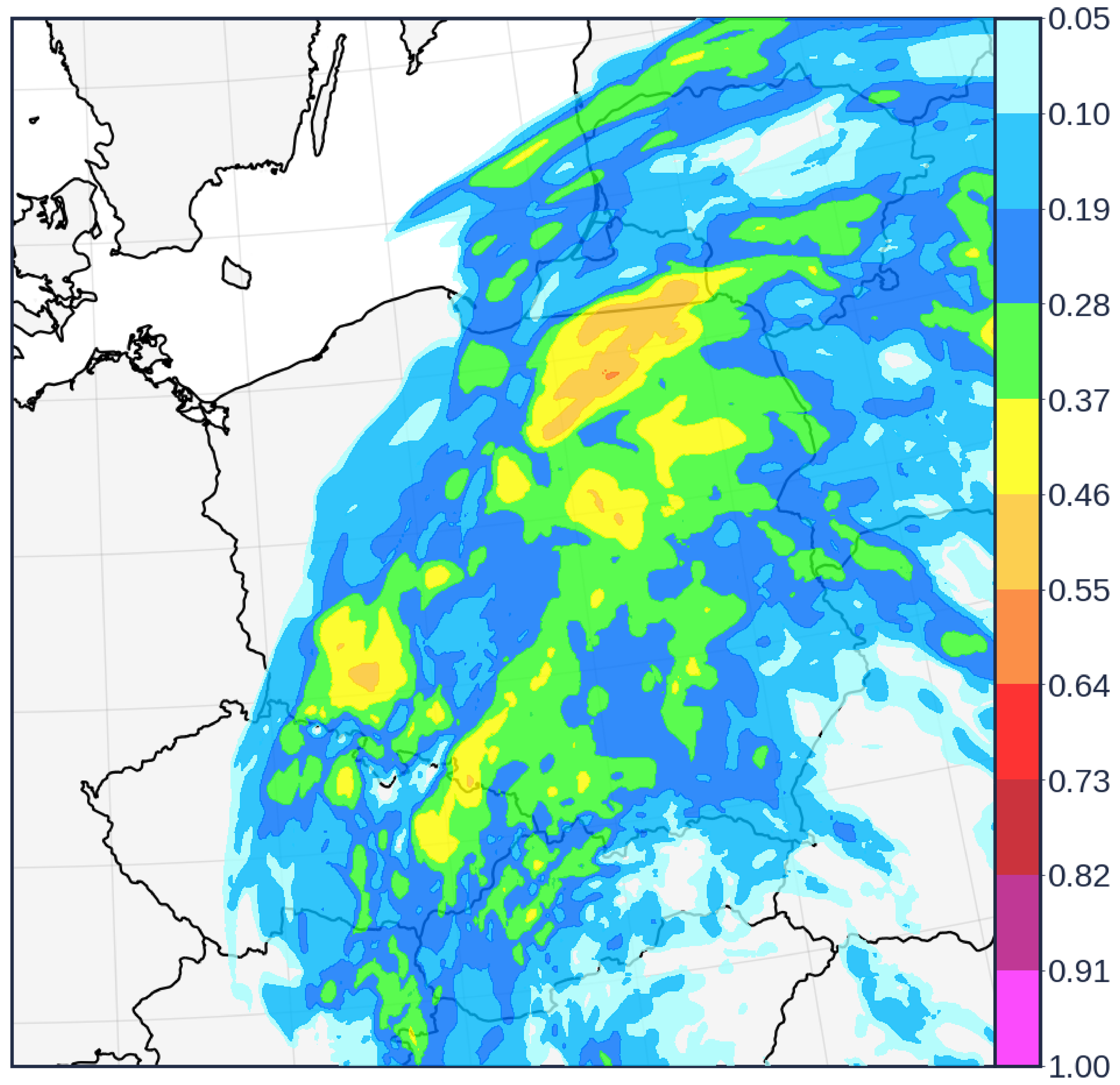
Outstanding Student & PhD candidate Presentation contest

Introduction

The 3D-Var method for data assimilation estimates atmospheric states by minimizing a cost function that measures the mismatch between model forecasts and observations, weighted by their error covariances. Standard implementations employ conjugate-gradient (CG) solvers. CG performs well for quadratic cost functions under Gaussian error assumptions, but in nonlinear or non-Gaussian settings, the overall minimization process may converge to suboptimal local minima. These conditions are characteristic of aviation turbulence assimilation, where measurements are spatially and temporally sparse and exhibit heterogeneous uncertainty.

This study assimilates Eddy Dissipation Rate forecasts from the COSMO time-lagged ensemble with turbulence observations derived from Mode-S EHS radar, as well as AMDAR and AIREP reports. The classical CG-based solver is compared with a hybrid metaheuristic framework, evaluating convergence characteristics and computational efficiency.

EDR – a measure of turbulence strength



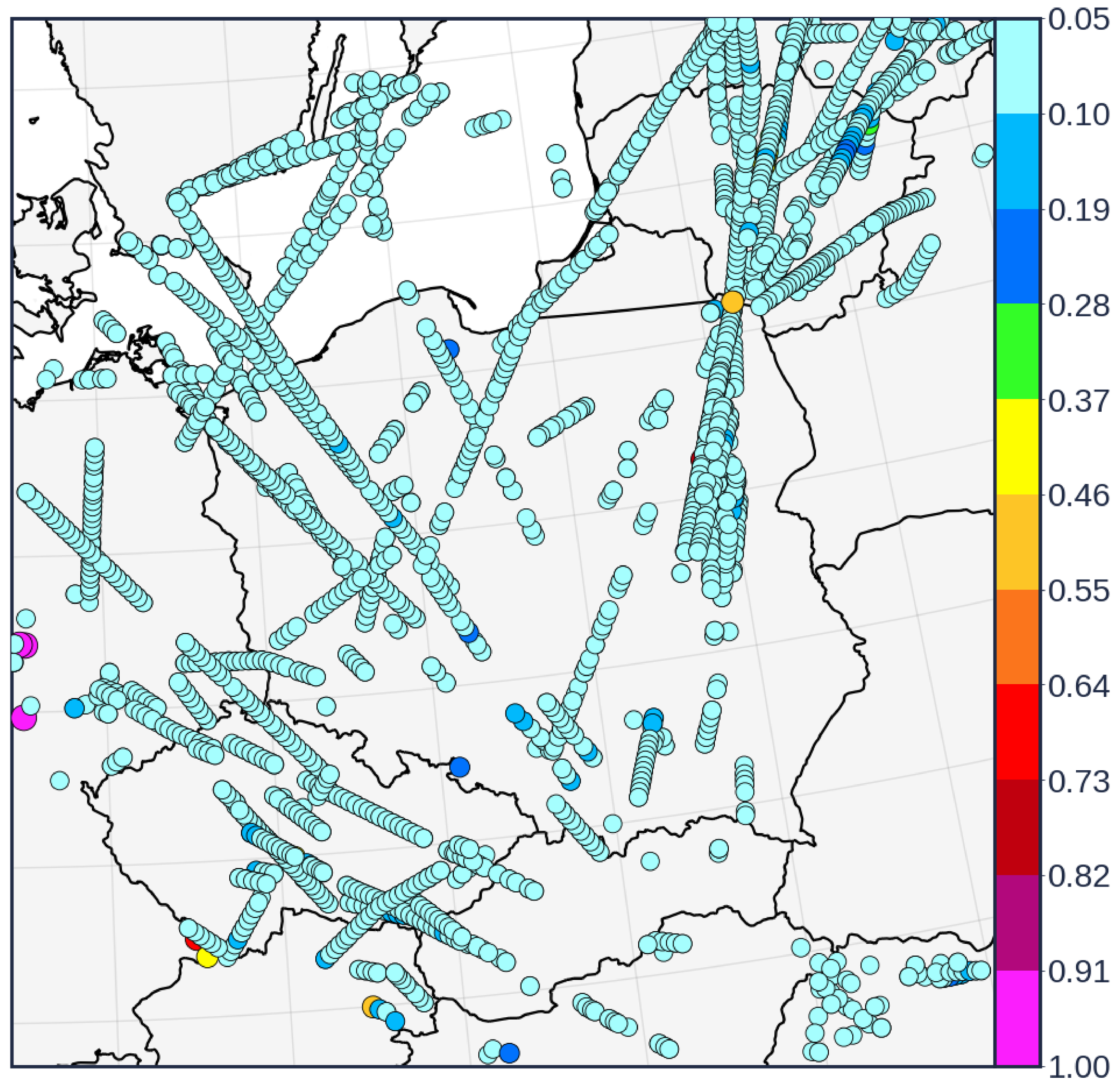
EDR ensemble forecast mean valid for 2025-11-26 00:00 UTC from the COSMO 2k8 TLE model. Color scale represents turbulence intensity.

- Eddy Dissipation Rate (EDR) [$m^{2/3}$ s^{-1}] is a standard metric used for turbulence forecasting and reporting
- EDR represents how quickly turbulent eddies in the atmosphere are broken down and dissipated
- COSMO 2k8 TLE model can produce EDR forecasts thanks to the Raschendorfer's turbulence parametrization scheme

Turbulence classification from aviationweather.gov (NOAA)

Aircraft weight class	EDR threshold			
	Light	Moderate	Severe	Extreme
Light	0.13	0.16	0.36	0.64
Medium	0.15	0.20	0.44	0.79
Heavy	0.17	0.24	0.54	0.96

Turbulence observations: Mode-S, AMDAR, AIREP



Turbulence observations from all data sources between 2025-11-25 22:00 UTC and 2025-11-26 02:00 UTC. Color scale represents turbulence intensity.

- Mode-S radar data records bring information about aircraft speed, heading and vertical rate; EDR can be calculated from a sequence of several consecutive clean Mode-S records within a short time window
- AMDAR reports provide in situ Derived Equivalent Vertical Gust measurements, which can be converted to EDR
- Pilots submit manual reports (AIREPs) with subjective turbulence assessments like "moderate" or "severe"
- All aircraft-based observations are sparse, confined to flight paths, and measured in heterogenous units

Metaheuristic algorithms for continuous optimization

- Population-based metaheuristics are widely used to solve continuous optimization problems
- They excel in black-box settings, as they do not require gradients or prior knowledge of the objective function
- These methods can outperform classical gradient approaches on multimodal, non-separable, or non-smooth landscapes
- Simple metaheuristic algorithms can be combined, exploiting the synergy between all the component algorithms
- The constrained Hybrid Metaheuristic (cHM) is a simple, yet very effective hybrid approach
- cHM alternates between a probing phase (evaluating several algorithms simultaneously) and a fit phase (continuing optimization with only the best-performing algorithm)
- This study uses a cHM integrating Particle Swarm Optimization, Genetic Algorithm, Ant Colony Optimization, Simulated Annealing, and Differential Evolution

3D-Var method for data assimilation

We start from the classical formulation of the cost function:

$$\mathcal{J}(x) = \tfrac{1}{2}\|x - x_b\|^2_{B^{-1}} + \tfrac{1}{2}\|y - H(x)\|^2_{R^{-1}}$$

The B matrix can be estimated using the NMC method:

$$B = \tfrac{1}{N-1}E^T E$$

where E is the centered error matrix calculated based on the differences between the forecasts valid for time t and $t + 6\text{h}$, and $N = 24{,}600$ is the number of samples in the E matrix.

Instead of calculating the B matrix directly, we approximate it as:

$$E \approx U_k \Sigma_k W_k{}^T \ \text{(truncated SVD)}$$

By defining $P = W_k \frac{\Sigma_k}{\sqrt{N-1}}$, we obtain:

$$B \approx W_k \left(\frac{\Sigma_k^2}{N-1}\right) W_k^T = PP^T$$

Using the control variable transform, we set $x = x_b + Pv$.

Observation errors are assumed to be uncorrelated, meaning R is a diagonal matrix, where the diagonal elements are the variances of each observation's error (σ_i^2). σ_i is set to 10% of the measured EDR for observation i, but not less than 0.05.

To mitigate the impact of data outliers, instead of using the standard quadratic L2 norm, we use the Huber norm:

$$\rho(e_i) = \begin{cases} \frac{1}{2}e_i^2, & |e_i| < 0.1 \\ |e_i| - 0.5, & |e_i| \geq 0.1 \end{cases}$$

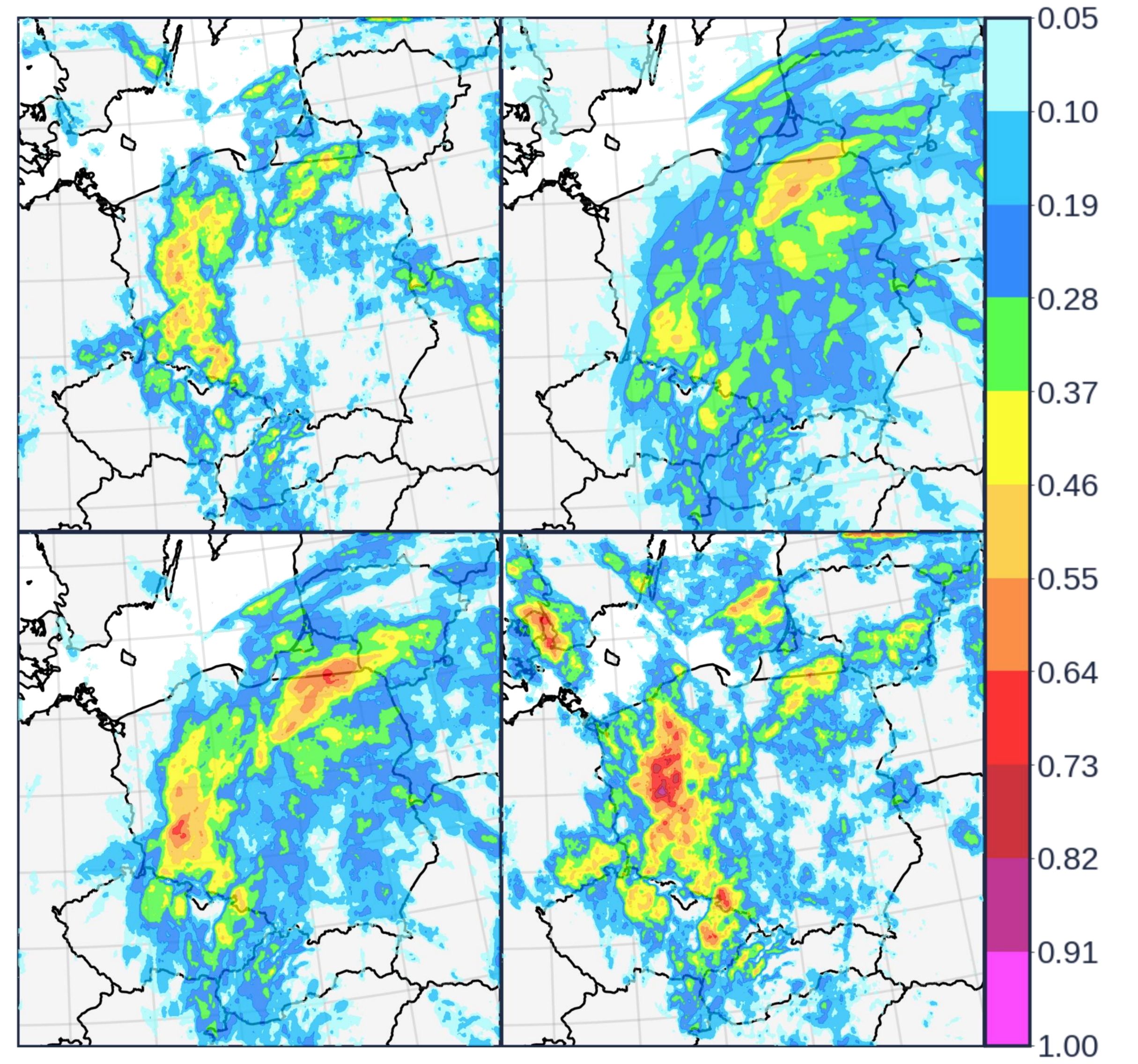
The final cost function to minimize is:

$$\mathcal{J}(v) = \tfrac{1}{2}v^T v + \sum_{i=1}^{N_{\text{obs}}} \rho\left(\frac{y_i - H(x_b + Pv)_i}{\sigma_i}\right)$$

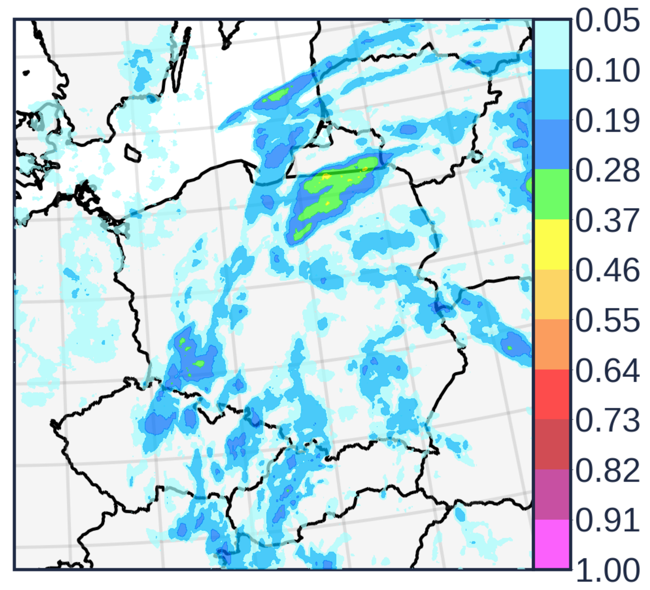
Results

Mean and standard deviation over 3360 folds (5 folds × 672 reanalysis construction tasks between 2025-11-24 00 UTC and 2025-12-21 23 UTC). Boldface indicates better value for each pair.

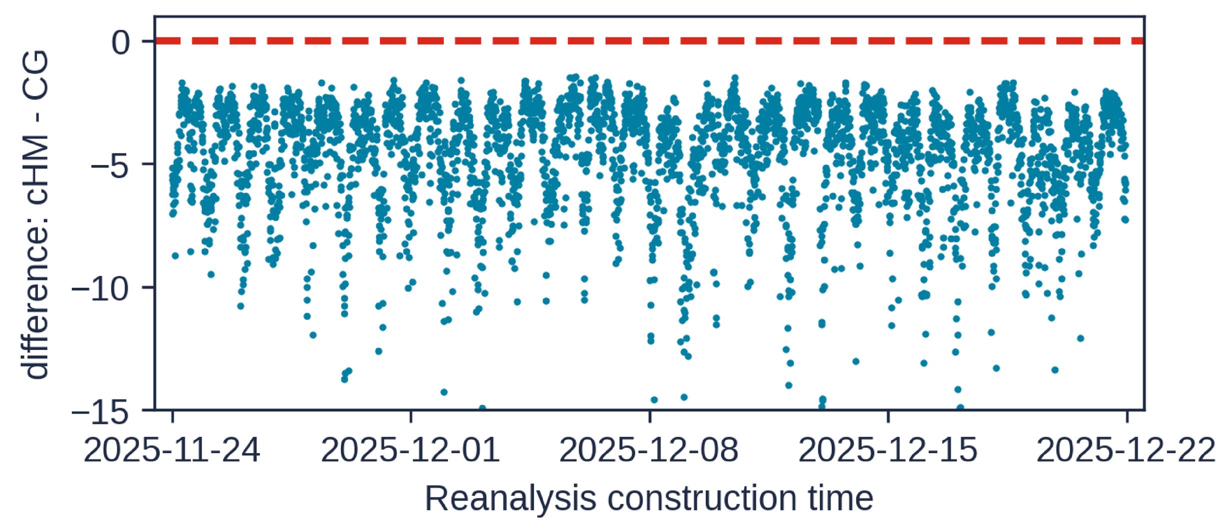
k	Method	RMSE	MAE	min $\mathcal{J}(v)$
2	CG	**.046054±.018999**	.021541±.006467	570.91±359.94
	cHM	.046097±.018997	**.021540±.006494**	**570.37±360.52**
5	CG	**.045744±.018816**	.021226±.006065	560.04±344.27
	cHM	.045786±.018824	**.021225±.006109**	**559.32±345.50**
10	CG	**.045438±.018641**	.020936±.005635	549.97±329.18
	cHM	.045468±.018642	**.020926±.005648**	**548.62±329.45**
25	CG	**.045040±.018547**	.020557±.005210	535.73±314.32
	cHM	.045053±.018542	**.020546±.005215**	**533.32±314.28**
50	CG	.044740±.018472	.020334±.004900	525.24±300.60
	cHM	**.044737±.018460**	**.020325±.004891**	**521.85±300.08**
100	CG	.044395±.018401	.020098±.004555	512.75±287.34
	cHM	**.044378±.018384**	**.020095±.004535**	**508.12±286.41**
climatology		.055532±.013556	.040632±.003132	n/a
first guess		.047334±.019865	.023423±.008712	n/a



Four solutions from the random initial population for the reanalysis for 2025-11-26 00:00 UTC. Color scale represents turbulence intensity.



Reanalysis obtained for 2025-11-26 00:00 UTC by optimization of the cost function (for $k = 100$) with cHM. Color scale represents turbulence intensity.



Differences in achieved cost function values between cHM and CG-based solver for all tasks for the truncation level $k = 100$

Conclusions

- For all truncation levels (k), cHM achieves lower cost function values than the CG-based solver
- For $k = 50$ and $k = 100$, cHM yields better RMSE and MAE against out-of-fold turbulence observations
- However, improvements in both the cost function and validation scores are marginal, with differences emerging only in the 5th or 6th decimal place

Acknowledgments

The research was partially supported by the National Science Centre, Poland, grant number 2023/49/B/ST6/01404, and project S-10/2026, „Aviation meteorology: development of methods and products for general aviation and drones".