The 8th International Conference on Deep Learning in Computational Physics,
June 19-21, 2024

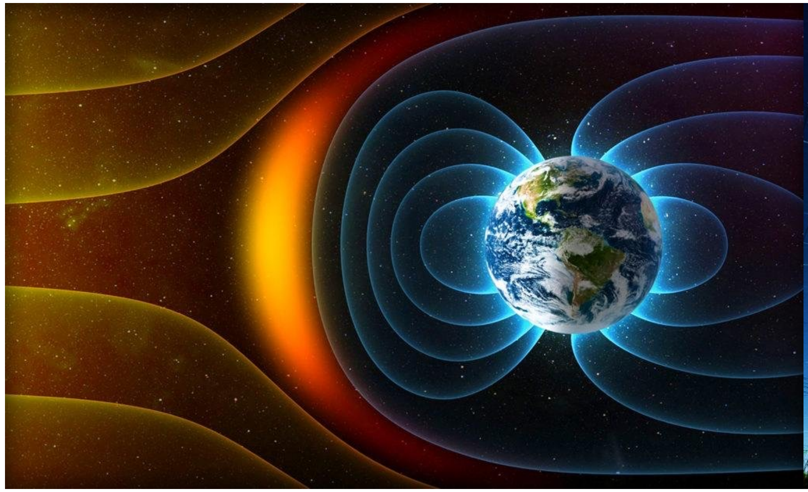
Comparative Analysis of the Procedures to Forecast the Kp Geomagnetic Index by Machine Learning

Gadzhiev I.M., Barinov O.G., Myagkova I.N., Dolenko S.A.
SINP MSU, Moscow

Исследование выполнено за счёт гранта Российского научного фонда, проект № 23-21-00237.

Prediction of Geomagnetic Disturbances

- Geomagnetic disturbances are one of the most important factors in space weather.
- They can cause disruption to radio communications, pipelines, power lines and electrical networks.
- Disturbance prediction can help to handle these problems.

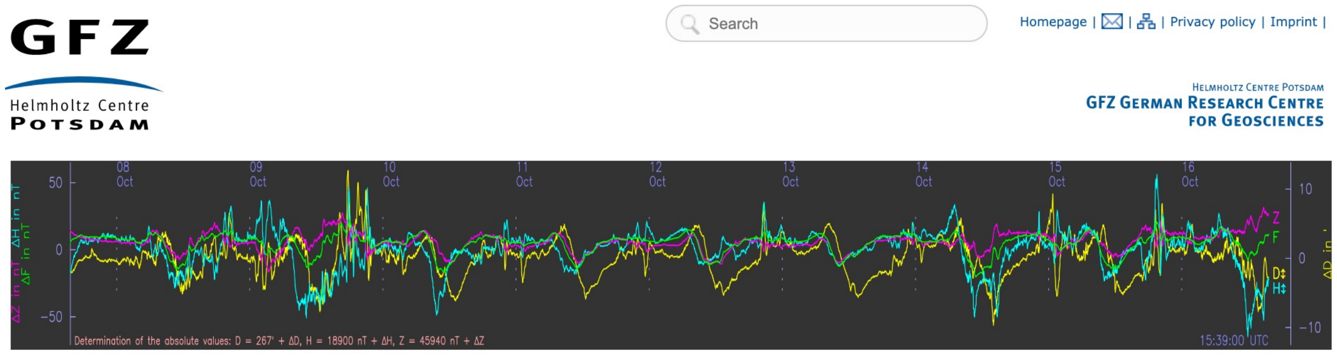


Planetary index Kp

One of the most used geomagnetic indices is the planetary index Kp ($Kp \in [0...9]$, with a step of 1/3)

Kp is the weighted average of K-indices obtained at 13 observatories. K-index is derived from the maximum fluctuations
of the horizontal components of the Earth's magnetic field,
observed on the magnetometer for 3 hours

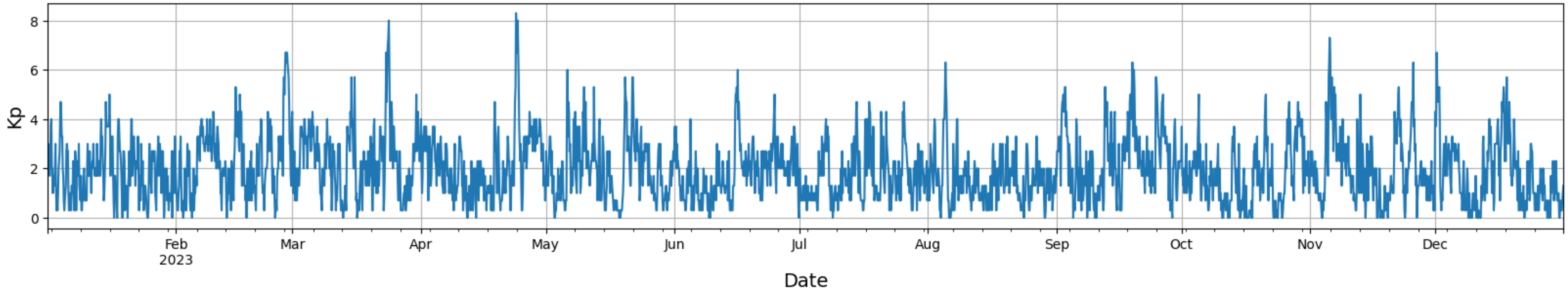
It has **3-hour frequency** (calculated at 03:00, 06:00, …)



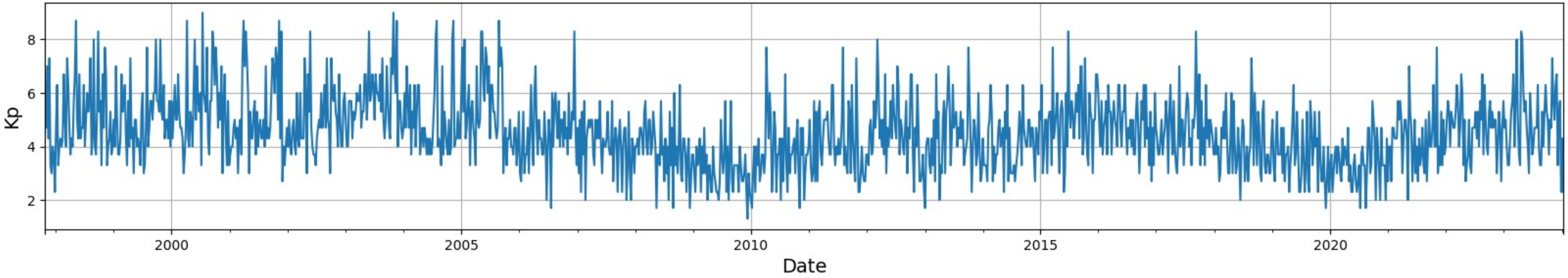
https://isdc.gfz-potsdam.de/kp-index/

Kp index dynamics

Dynamics for 2023 year



Weekly maximums since 1997



https://isdc.gfz-potsdam.de/kp-index/

Data for Kp forecasting

Feature	Description
Kp*10 (previous)	Index Kp 3-hour [1]
Dst	Index Dst [2]
B_x	x component of the magnetic field
B_gsm_y	x component of the magnetic field (GSM)
B_gsm_z	y component of the magnetic field (GSM)
B_magn	Magnetic field module
SW_spd	Solar wind speed at the Lagrange point L1
H_den_SWP	Solar wind density at the Lagrange point L1
daySin	sin(2*π*[Day of a year]/365)
hourSin	sin(2*π*[Hour of a day]/24)
dayCos	cos(2*π*[Day of a year]/365)
hourCos	cos(2*π*[Hour of a day]/24)
Trr_SWP	Solar wind temperature at the Lagrange point L1

The goal of the study

Given some previous data, the goal is to predict the next 8 values of Kp index

Kp index at horizon i = next i-th value of the Kp index.
For each horizon **we train a single model*.**

Because the Kp-index has 3 hours frequency, the first statement expands to:

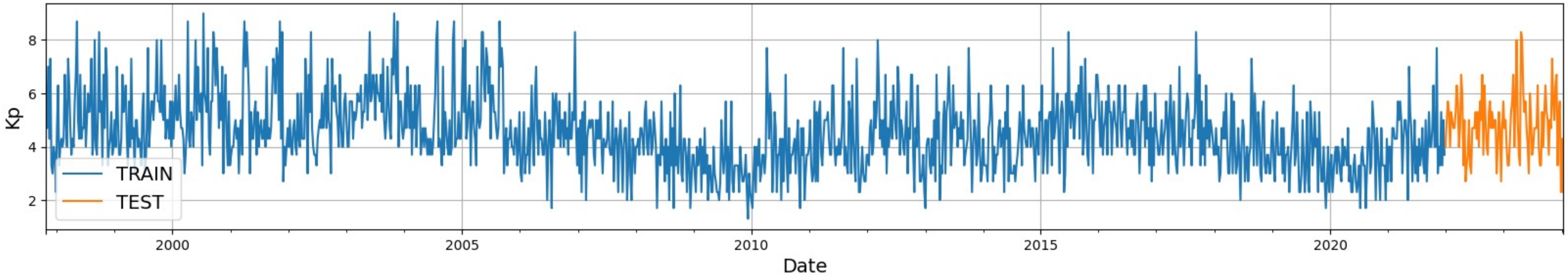
- At 00:00, 03:00, 06:00 … (hereinafter **T0** hours), **predict Kp index from 3 to 24 hours ahead inclusively with a 3-hour step;**
- At 02:00, 05:00, 08:00, … (hereinafter **T1** hours), **predict Kp index from 1 to 22 hours ahead inclusively with a 3-hour step**;
- At 01:00, 04:00, 07:00, … (hereinafter **T2** hours), **predict Kp index from 2 to 23 hours ahead inclusively with a 3-hour step.**

*See P.S

The quality assessment

The range of values of all the variables for 2022-2024 was used as a test set for assessing quality of models.

Weekly maximums of Kp divided into training and test sets:



The quality metrics

We use **Root Mean Squared Error (RMSE) as a primary metric** to assess quality of regression models.

Some additional metrics include:

- Mean Absolute Error (MAE);
- Accuracy – because Kp-index is a discrete index, we can round prediction of a model to the nearest value of Kp-index and measure classification accuracy of the rounded prediction.

Model setup

- Linear model: Ridge
- Gradient boosing: LightGBM implementation (LightGBM) and CatBoost implementation (CatBoost)
  - Max Tree Depth = 3, max number of trees 10000, early stopping with 50 iterations, learning rate 0.1.
- Multi-layer Perceptron
  - MLP1 – 1 hidden layer with 128 neurons.
  - MLP2 - 2 hidden layers with 256 and 128 neurons.
  - 200 epochs, early stopping with 30 iterations, learning rate 0.03 with reducing on validation loss plateau after 30 iterations, Adam optimizer.
- Trivial model = Inertial forecast (next value of Kp = previous)

Feature preprocessing

The forecasting of next Kp-index values should probably require a history of all the parameters before the prediction moment.

To take this history into the account **we use delay embedding technique** – vector of variables for prediction moment $x(t_i)$ is concatenated with vectors for N previous moments $\{x(t_i), x(t_{i-1}),\dots, x(t_{i-N})\}$ and then fed into a model. **N is the depth** of the delay embedding.

We test delay embedding of **0 (no delay) to 24 hours with step 3** (period of Kp)

Baseline, depth 0

Green highlights the best model for each horizon.
All results should be compared with inertial forecast at the bottom

Horizons 1-3

Horizon	1			2			3		
	Accuracy	R2	RMSE	Accuracy	R2	RMSE	Accuracy	R2	RMSE
Catboost	22,9%	0,748	6,372	18,4%	0,572	8,312	**15,1%**	**0,395**	**9,88**
LightGBM	22,7%	0,747	6,394	18,7%	0,569	8,335	15,0%	0,391	9,908
MLP1	22,5%	0,75	6,347	18,3%	0,57	8,324	15,2%	0,392	9,901
MLP2	**22,7%**	**0,753**	**6,312**	**18,4%**	**0,573**	**8,297**	14,9%	0,386	9,954
Ridge	20,2%	0,701	6,949	15,9%	0,518	8,818	13,6%	0,356	10,195
Inertial	18,0%	0,495	9,031	13,0%	0,163	11,62	12,3%	-0,023	12,848

Horizons 4-6

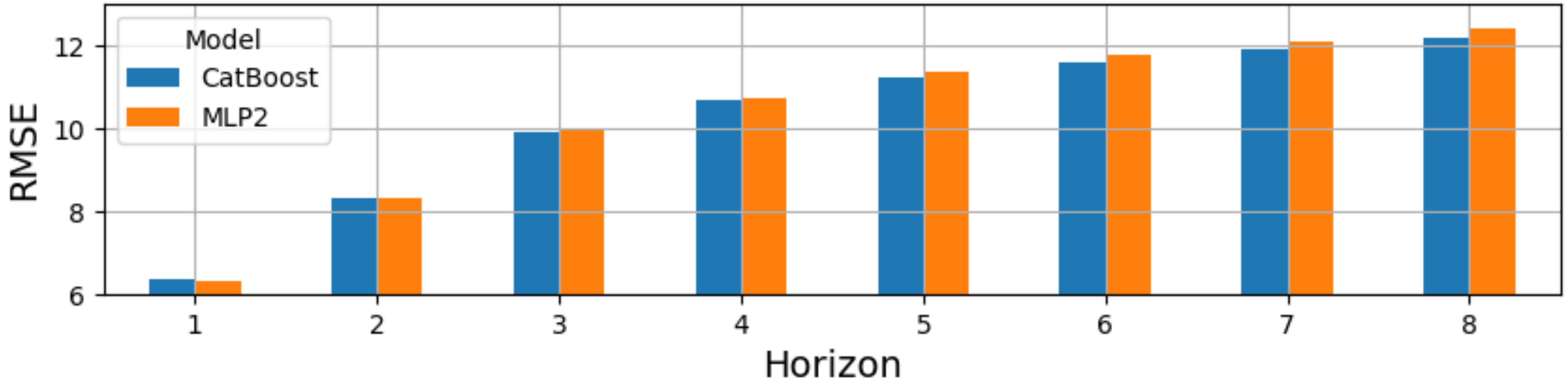
Horizon	4			5			6		
	Accuracy	R2	RMSE	Accuracy	R2	RMSE	Accuracy	R2	RMSE
Catboost	**14,0%**	**0,293**	**10,677**	**12,3%**	**0,222**	**11,204**	**11,8%**	**0,166**	**11,6**
LightGBM	13,7%	0,289	10,712	12,5%	0,215	11,254	11,8%	0,161	11,637
MLP1	13,8%	0,285	10,739	12,9%	0,21	11,287	12,1%	0,147	11,733
MLP2	14,6%	0,288	10,715	13,0%	0,202	11,347	12,3%	0,141	11,77
Ridge	12,8%	0,256	10,959	12,1%	0,18	11,504	11,7%	0,124	11,888
Inertial	12,0%	-0,162	13,693	10,8%	-0,281	14,379	10,4%	-0,384	14,943

Horizons 7-8

Horizon	7			8		
	Accuracy	R2	RMSE	Accuracy	R2	RMSE
Catboost	11,6%	0,122	11,901	11,8%	0,085	12,147
LightGBM	11,4%	0,114	11,958	11,6%	0,078	12,194
MLP1	11,7%	0,109	11,987	11,6%	0,075	12,215
MLP2	11,7%	0,095	12,082	12,0%	0,048	12,393
Ridge	11,3%	0,078	12,192	11,2%	0,046	12,404
Inertial	10,6%	-0,473	15,415	10,5%	-0,543	15,775

Baseline, depth 0, interpretation

- MLP with 2 hidden layers wins at first two horizons, CatBoost wins at other horizons. Comparison plot is below.
- For the horizon 1 best **RMSE 6.312** (~2 units of Kp), **Accuracy 22.7%** (Compared to RMSE 9.031 and Accuracy 18% of inertial forecast);
- For the horizon 8 best **RMSE 12.147, Accuracy 11.8%** (Compared to 15.775 and 10.5% for inertial)

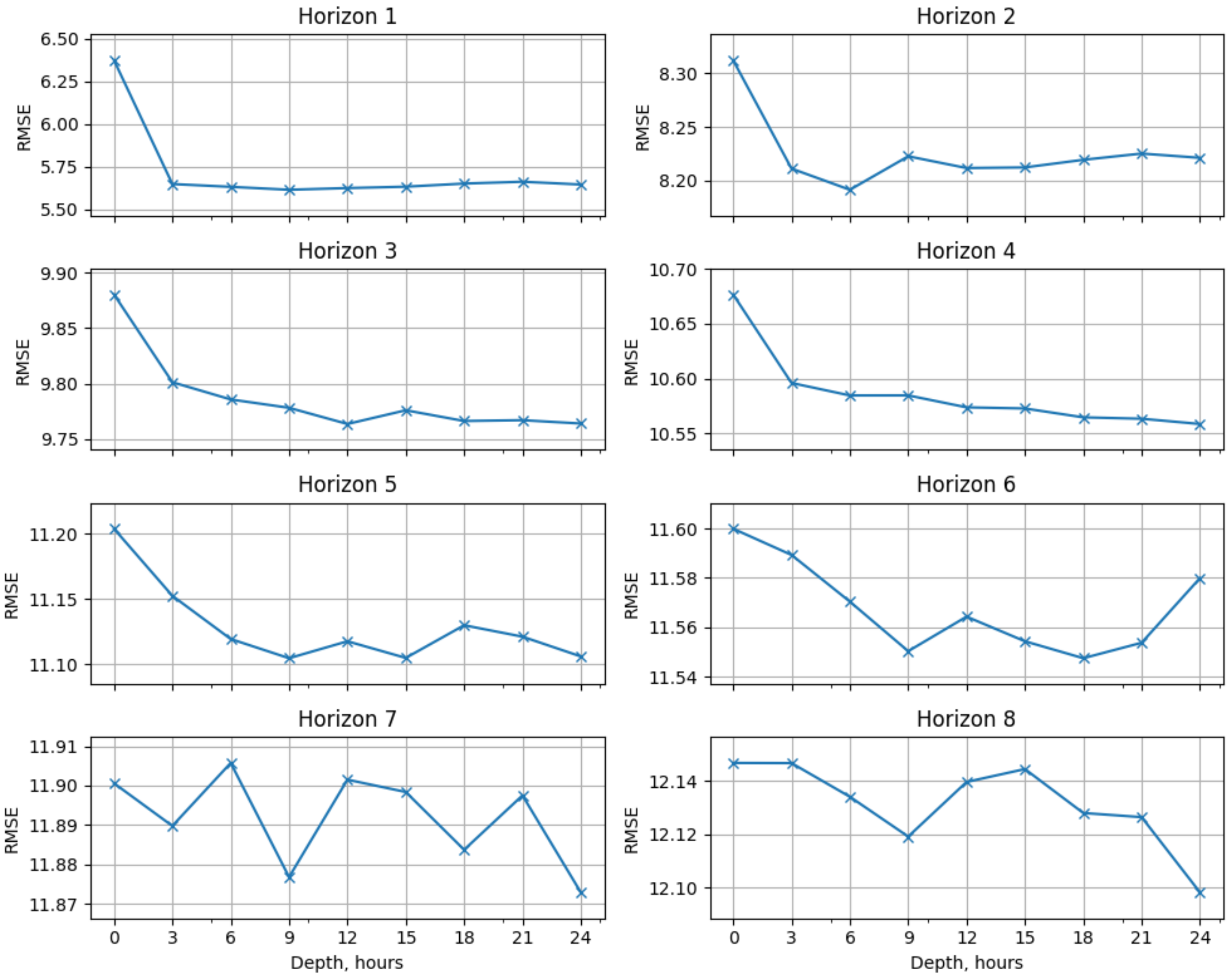


Increasing depth of delay embedding

At the left –
the dependence of RMSE
on depth
for CatBoost

For nearest horizons –
lower depth is needed

For farthest – higher.



Comparing with the baseline

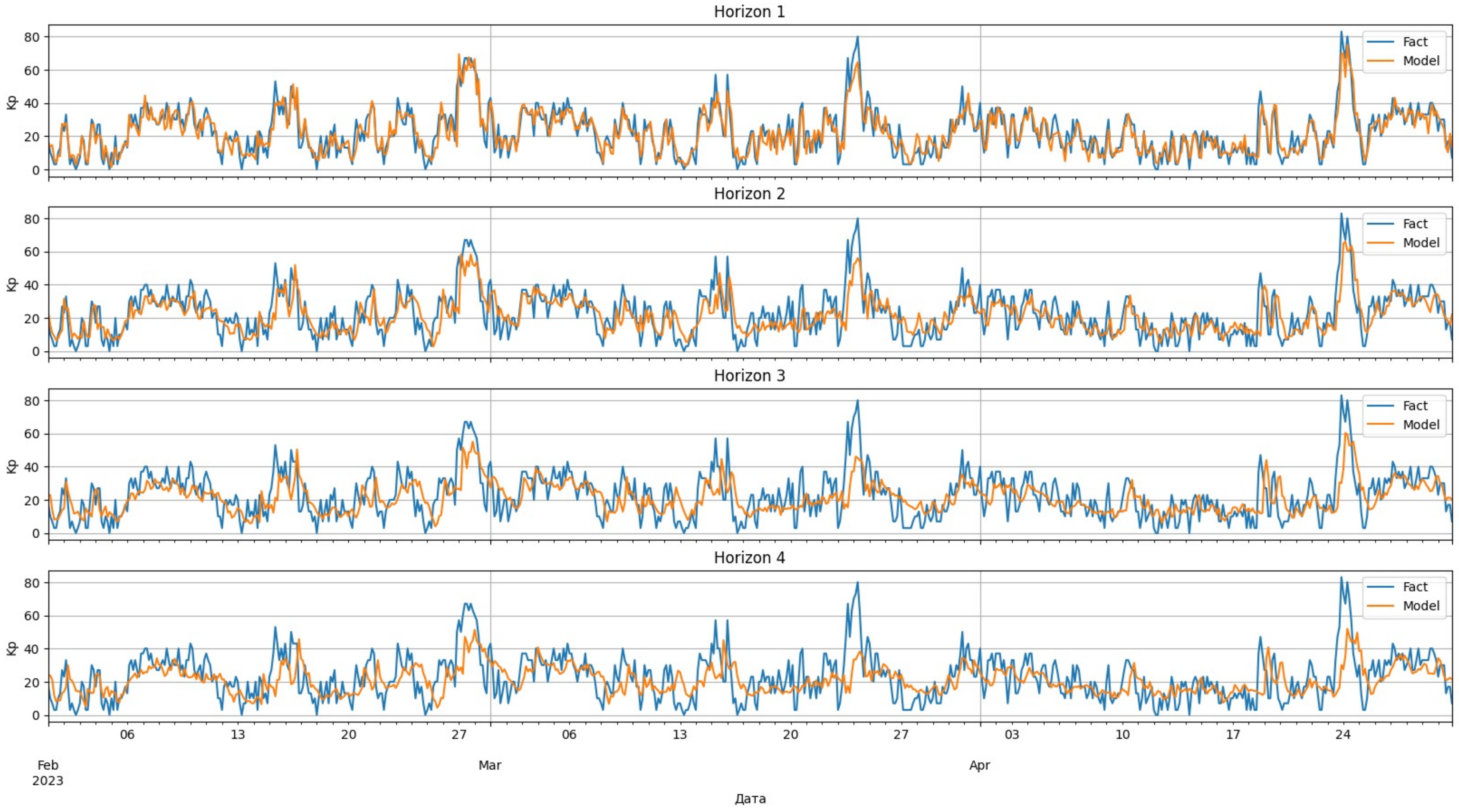
Table below shows best model trained for each horizon together with the depth.

- For horizon 1 we obtained **RMSE 5.615 and Accuracy 26.7%** (RMSE 6.312, Accuracy 22.7% for the baseline);
- For horizon 8 – **RMSE 12.098 and Accuracy 12.1%** (RMSE 12.147, Accuracy 11.8% for the baseline).

Horizon	Best Model	Accuracy	MSE	R2
1	CatBoost 9H	26,7%	5,616	0,805
2	CatBoost 6H	18,7%	8,191	0,584
3	CatBoost 12H	15,1%	9,764	0,409
4	CatBoost 24H	13,6%	10,559	0,31
5	CatBoost 9H	12,8%	11,105	0,236
6	CatBoost 18H	12,3%	11,548	0,174
7	CatBoost 24H	11,7%	11,873	0,127
8	CatBoost 24H	12,1%	12,098	0,093

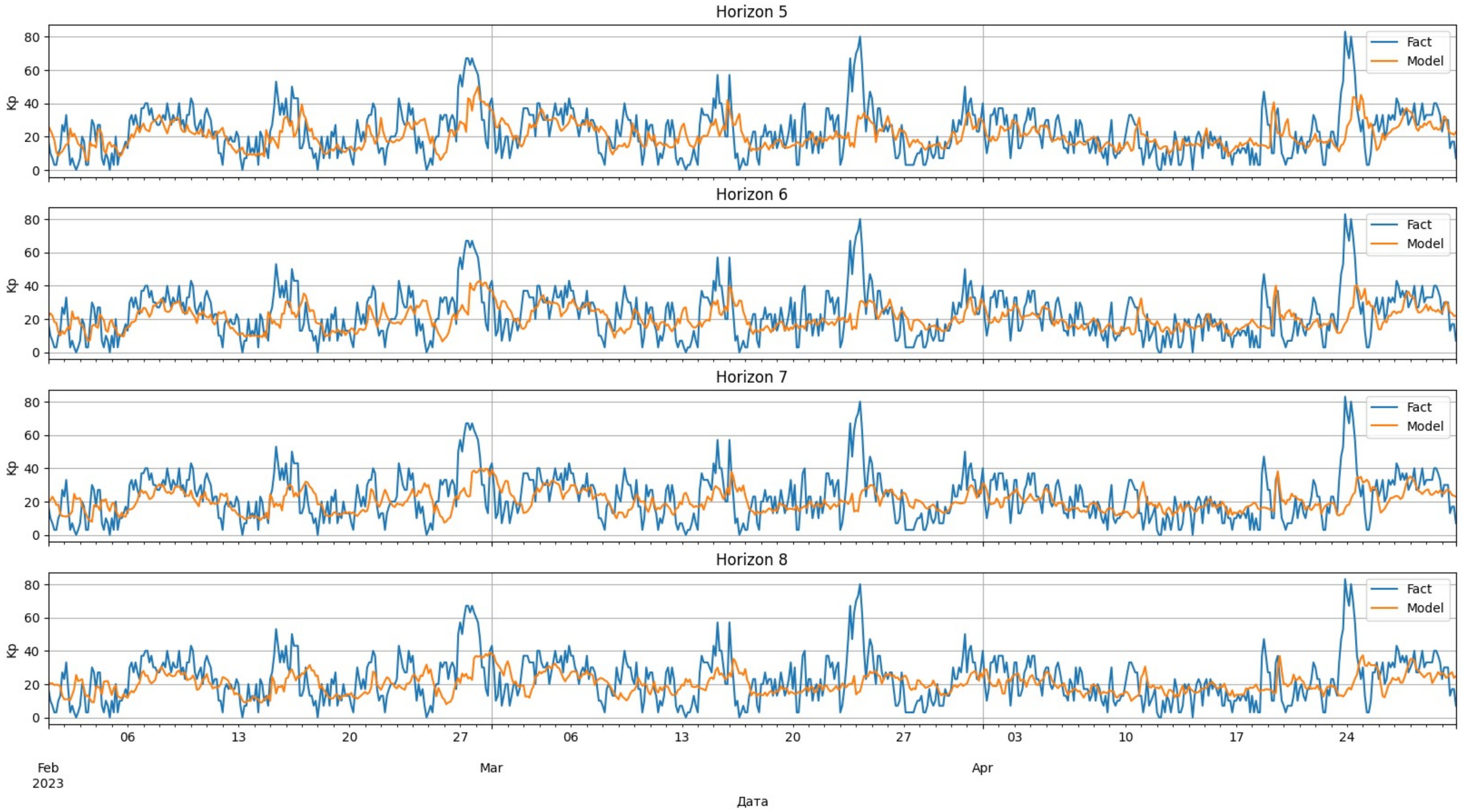
Best model demo (1-4 horizons) Feb-Apr 2023

CatBoost with delay embedding 9 hours (best at 1st horizon)



Best model demo (5-8 horizons) Feb-Apr 2023

CatBoost with delay embedding 9 hours (best at 1st horizon)



Feature importances for the best model

CatBoost with delay embedding 9 hours, feature importances for the 1st horizons. X-axis shows the feature, y-axis – the delay.

The most important features seem to be **Kp itself (preceding value), B_gsm_z and its lags, B_magn, SW_spd and Trr_SWP**

	Kp	doySin	hourSin	doyCos	hourCos	Dst	B_x	B_gsm_y	B_gsm_z	B_magn	SW_spd	H_den_SWP	Trr_SWP
0	41.22	0.21	0.06	0.07	0.04	0.71	0.82	0.73	10.48	8.22	9.64	3.33	1.09
1		0.04	0.06	0.03	0.02	0.11	0.43	0.35	7.20	0.81	0.90	0.20	0.11
2		0.08	0.06	0.03	0.25	0.07	0.22	0.21	2.73	0.50	0.11	0.06	0.10
3	0.53	0.04	0.09	0.03	0.02	0.06	0.06	0.11	0.74	0.11	0.06	0.07	0.07
4		0.07	0.06	0.03	0.09	0.04	0.03	0.05	0.19	0.06	0.03	0.07	0.05
5		0.07	0.04	0.03	0.20	0.05	0.04	0.05	0.53	0.09	0.04	0.09	0.04
6	0.22	0.03	0.00	0.03	0.01	0.05	0.04	0.04	0.14	0.05	0.04	0.04	0.05
7		0.08	0.06	0.03	0.00	0.03	0.03	0.03	0.05	0.05	0.03	0.07	0.06
8		0.08	0.25	0.03	0.00	0.04	0.03	0.02	0.04	0.03	0.05	0.07	0.07
9	0.21	0.13	0.01	0.08	0.02	0.07	0.04	0.04	0.07	0.04	0.12	0.19	0.08

Conclusions

- We tested different ML models against the task of predicting next 8 Kp-index values;
- We showed, that inertial model forecast could be beaten using just current value. For horizon 1 we obtained **RMSE 6.312,** for horizon 8 **- 12.147.** The best models are CatBoost and perceptron with 2 hidden layers;
- We showed, that these results could be outperformed if history is taken into the account. Using CatBoost with delay embedding we obtained **RMSE 5.615** for horizon 1**, RMSE 12.098** for horizon 8;
- The most important features for the forecast are **Kp itself, B_gsm_z and its lags, B_magn, SW_spd and Trr_SWP.**

Thank you for your attention!

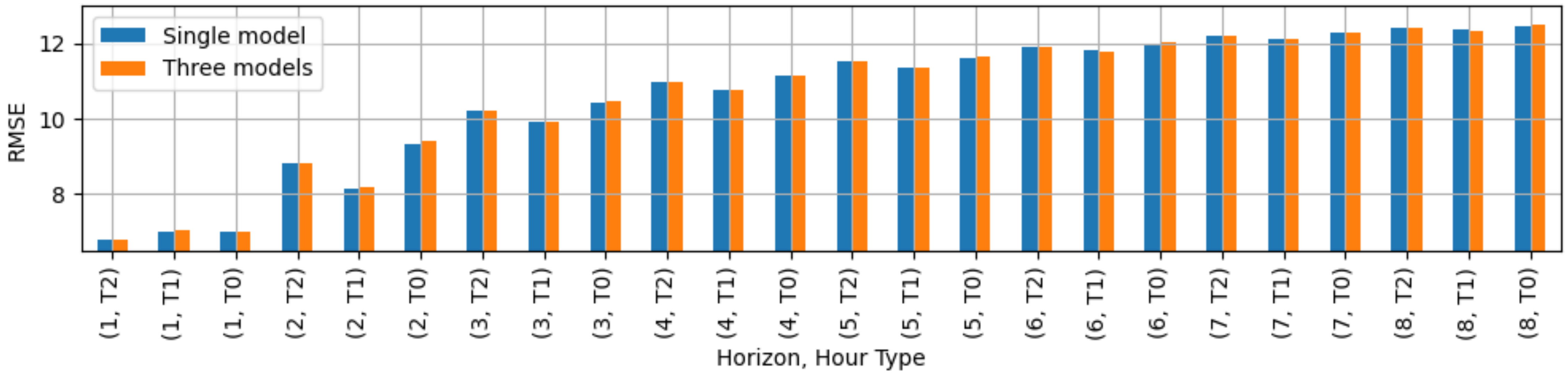
P.S.

Three models for each Hour Type or single model?

- There are three different positions, from which we can forecast the Kp – the T0, T1, T2 hours.
  At each position **there is a different number of hours till the next Kp-measurement**;
- We already have 8 models for each of 8 horizons (8 next Kps).
  The question is – should we train a separate model for each hour type T0, T1, T2?

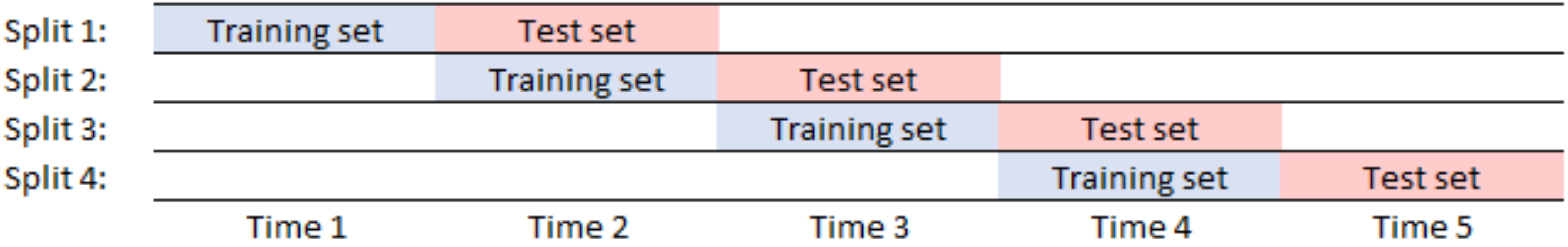
Single model is more convenient

- We train different **Ridge** model for each hour type
(3 hour types and 8 horizons = 24 models)
and single **Ridge** model for all types (8 horizons = 8 models);
- Plot below displays comparison of RMSE for each horizon and hour type for three models and single model.
It shows, that **this has little or no effect on the metric.**



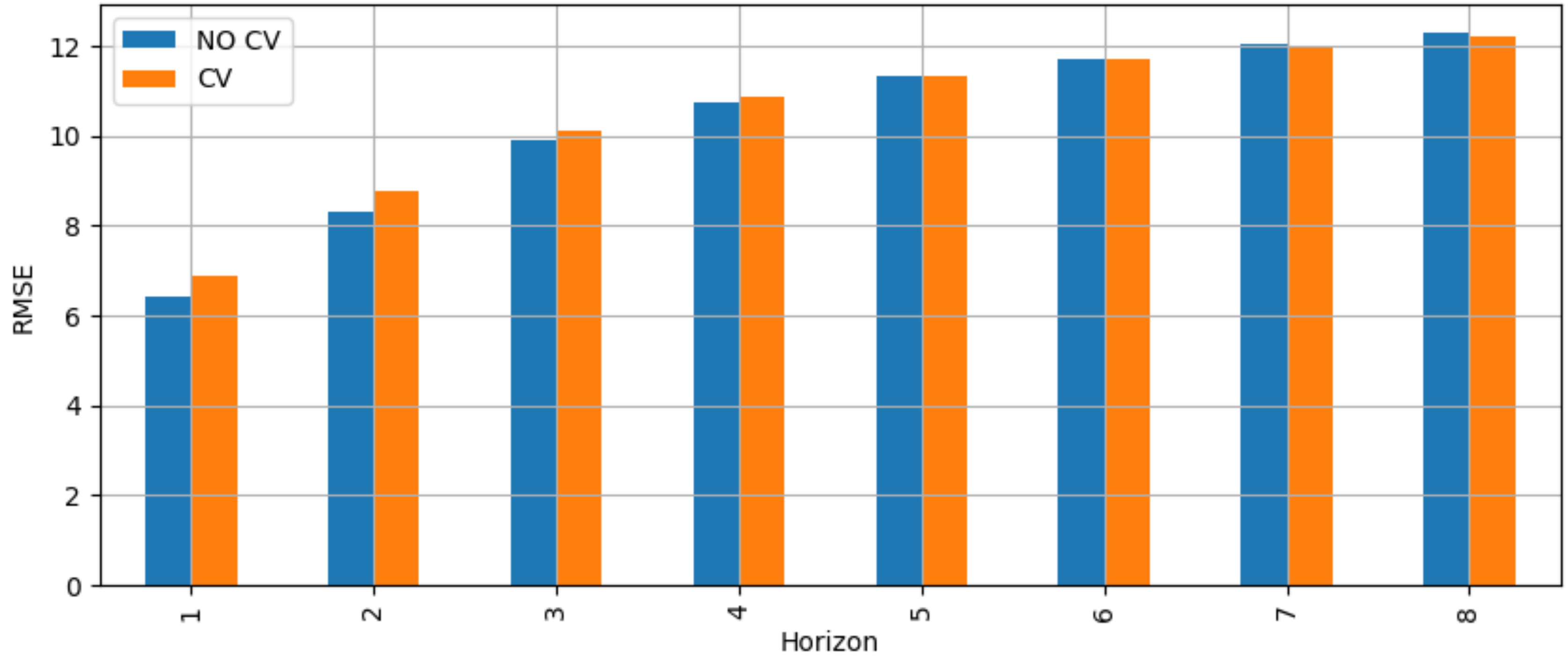
To use CV or not?

We also tested cross-validation technique to tune hyperparameters of GB and Ridge. CV is done with time-series split, as shown below. At each split model is trained and tested,
then RMSE is averaged over all splits.



CV does not seem to work in this case

Comparison of default models and model with hyperparameters tuned with this kind of CV shows that CV does not improve RMSE of a model in this task



How to choose validation set?

- Validation set could be used for early stopping for Gradient Boosting (number of trees) and MLP (number of epochs)
- **Option 1 – Validation set from 2020 to 2021.**
  - Pros – Kp-index has domain drifts, we can simulate that. Also, this period is close to test period.
  - Cons – models may overfit to this particular period (and this period could be an anomaly)
- **Option 2 – Randomly sample validation set from 1997 to 2021**
  - Pros – patterns from different periods in validation set to make it more robust;
  - Cons – not testing against domain drift.

Option 1 seems to be crucial for GB

- Results for MLP show, that there is no much difference for models with 1st or 2nd option
- Results for **LightGBM and CatBoost show, that domain drift simulation is crucial.** Plot below is RMSE for Option 1 (Separate) and Option 2 (Random) for each horizon. **Option 1 is better.**