Машинное обучение для статистической детализации характеристик пространственного распределения осадков в Московском регионе

Ярынич Юлия [1,2], Варенцов Михаил [1,2,3], Криницкий Михаил [4,5,1], Степаненко Виктор [1,2]

1 МГУ М.В. Ломоносова, Научно-исследовательский вычислительный центр

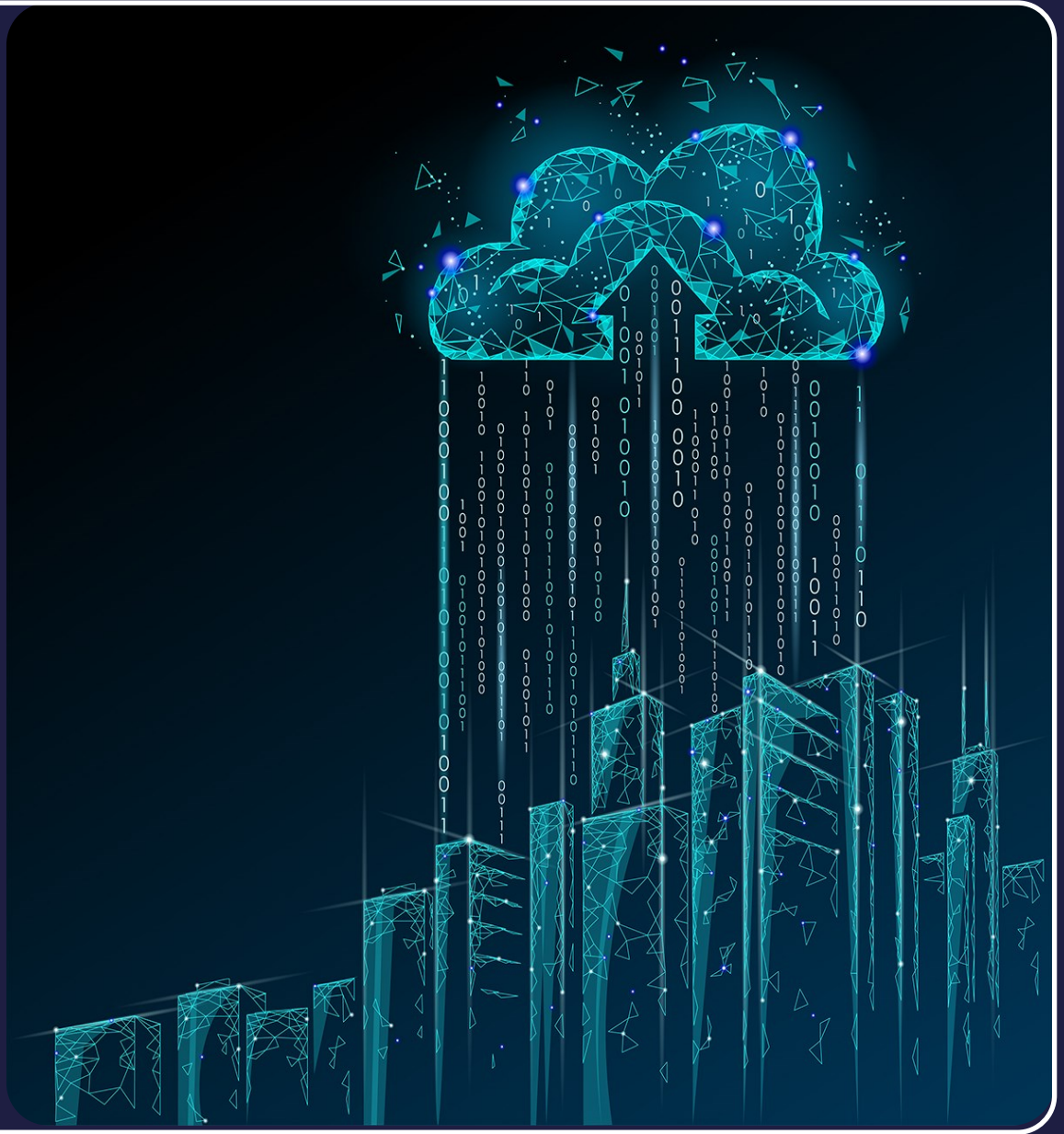
2 Институт физики атмосферы имени А.М. Обухова РАН

3 Гидрометеорологический научно-исследовательский центр РФ

4 Институт океанологии им. П.П. Ширшова РАН

5 Московский физико-технический институт

e-mail: julia.yarinich@srcc.msu.ru



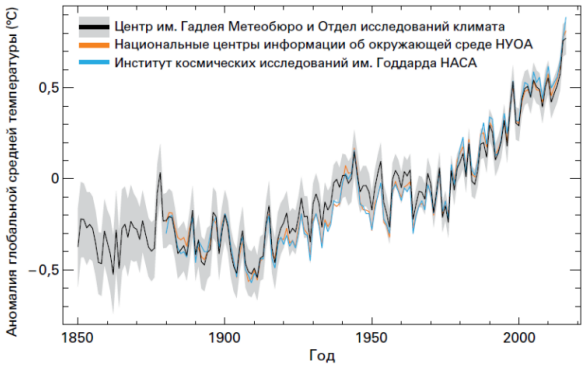
03 июля 2025

Актуальность:

Изменение климата

Повышение глобальной (приземной) **температуры**

Выше интенсивность и частота **экстремальных ливней**

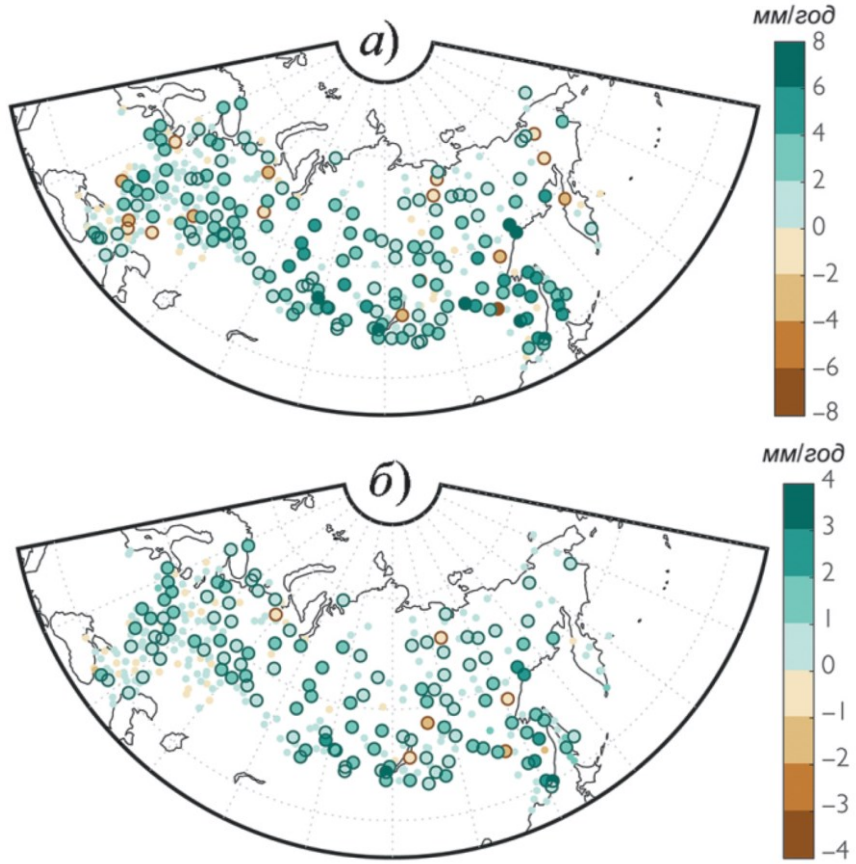


Средняя сезонная сумма осадков

Средняя сезонная сумма осадков **95 процентиля**

Тренд по данным наблюдений

за тёплый сезон (апрель – октябрь) 1966 - 2020



Акентьева Е. М. и др. Доклад о климатических рисках на территории Российской Федерации. – 2017.

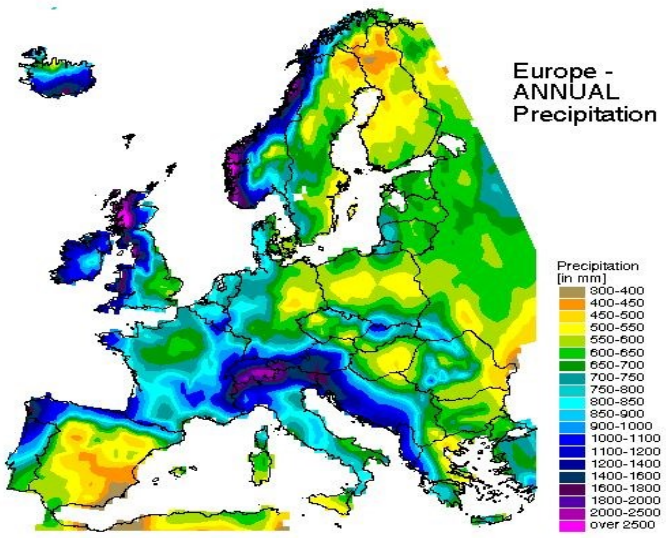
Chernokulsky A. V. et al. Atmospheric severe convective events in Russia: Changes observed from different data //Russian Meteorology and Hydrology. – 2022. – Т. 47. – №. 5. – С. 343-354.

Проблема:

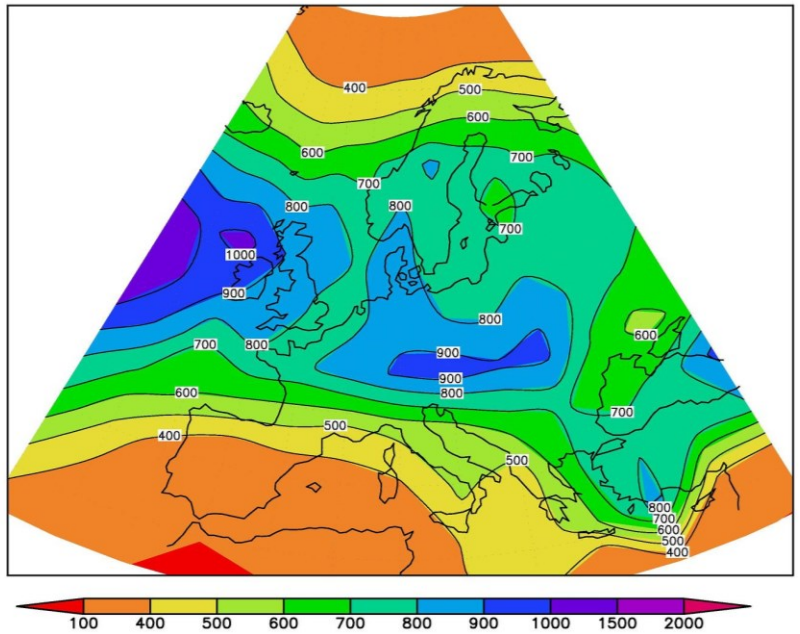
прогноз осадков с высоким пространственным разрешением

HIRESMIP – результаты моделирования осадков в Европе
Обзор модели ИВМ РАН

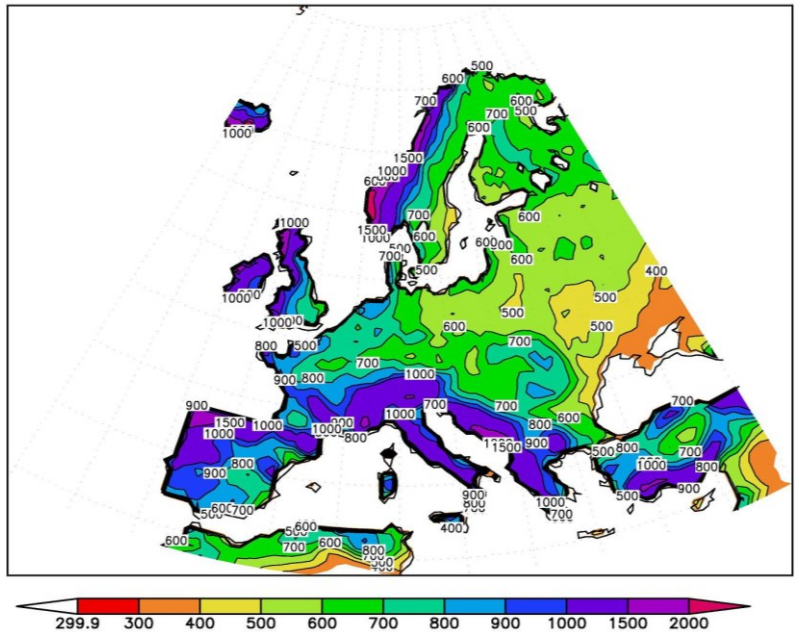
Реанализ («реальность»)



Модель 5x4 градуса

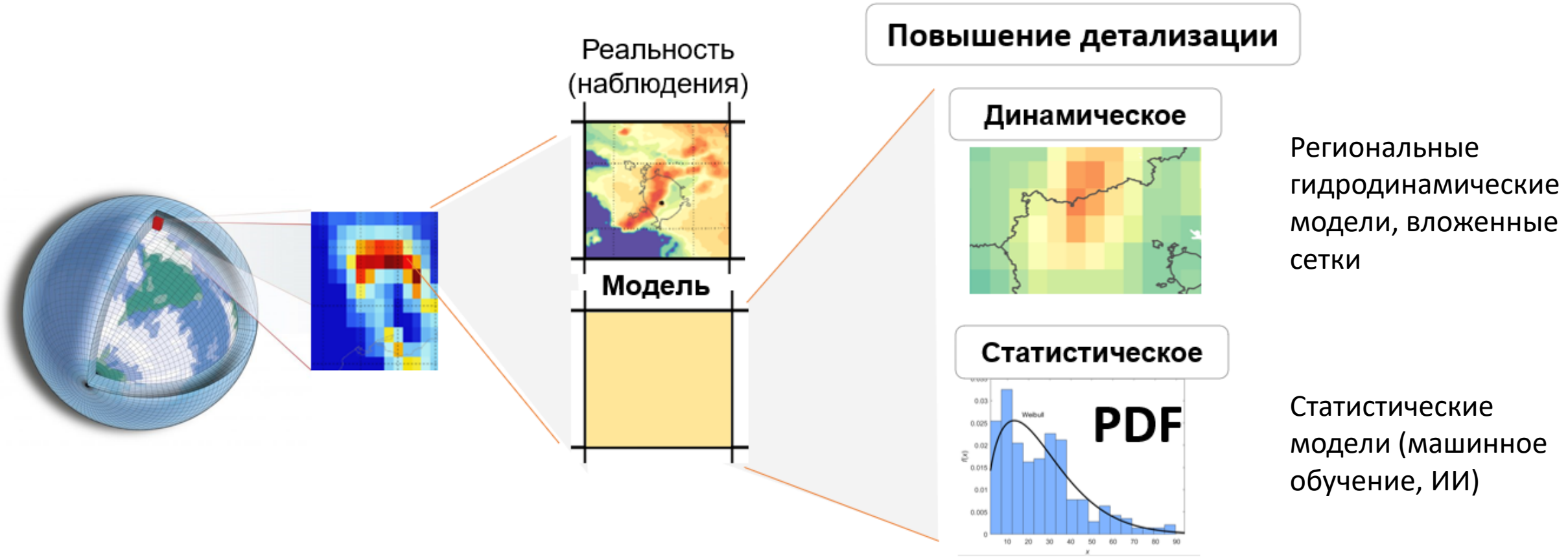


Модель 0.66x0.5 градуса



Проблема:

прогноз осадков с высоким пространственным разрешением



Региональные гидродинамические модели, вложенные сетки

Статистические модели (машинное обучение, ИИ)

Актуальность и обзор проблемы:

обоснованность применения статистических методов

Схематичное описание методики статистического даунскейлинга. [Dierickx, 2019]

- Значения в точках
- Менее вычислительно затратно
- Применяется к глобальным и региональным моделям

- Стационарность статистических зависимостей
- Длинные ряды данных наблюдений
- **Выбор предикторов**

Методика и материалы:

Постановка задачи

Целевая переменная

Признаки (предикторы)

Модели

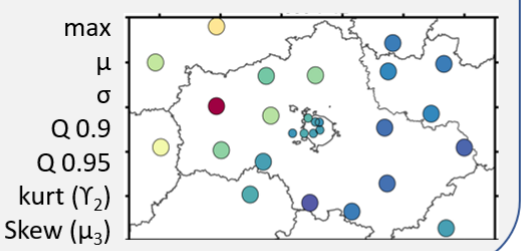
Разработка **технологии, позволяющей получать вероятностные характеристики интенсивных осадков** на основе поиска **взаимосвязей между данными об осадках** на метеостанциях и **крупномасштабными предикторами**, описывающими состояние атмосферы, наиболее благоприятствующее формированию мощных конвективных систем и связанных с ними интенсивных и экстремальных осадков

Целевая переменная (предиктант)

Данные наблюдений

(суточные суммы осадков и статистики пространственного распределения за каждые сутки)

max
μ
σ
Q 0.9
Q 0.95
kurt (Υ_2)
Skew (μ_3)



Обучение модели

$$y = f(X, g) + \varepsilon$$

Предиктор

Данные реанализа ERA5

35 физически обоснованных предикторов осадков

Модельные осадки

Стратификация

Условия для ДКС*

Влажность

Крупномасштабная циркуляция

x 35

*ДКС – долгоживущие конвективные системы

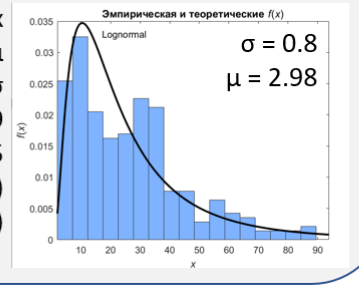
Как внутри ячейки сетки распределены осадки?

Применение модели

Результат детализации

Функция и характеристики подсеточного распределения суточных сумм осадков по пространству

max
μ
σ
Q 0.9
Q 0.95
kurt (Υ_2)
Skew (μ_3)



Методика и материалы:

Постановка задачи

Характеристики **пространственного распределения** суточных сумм осадков по данным наблюдений

Целевая переменная

Параметры эмпирического распределения

среднее

максимум

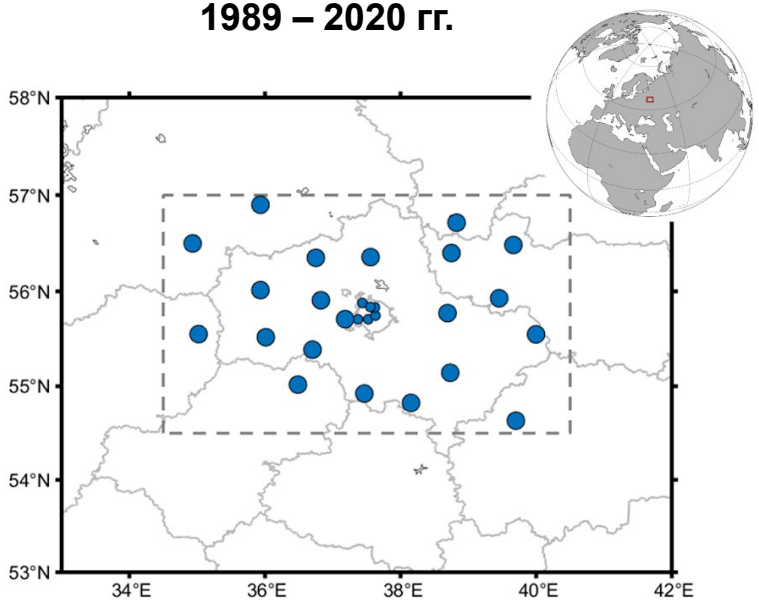
СКО

квантиль 0.9

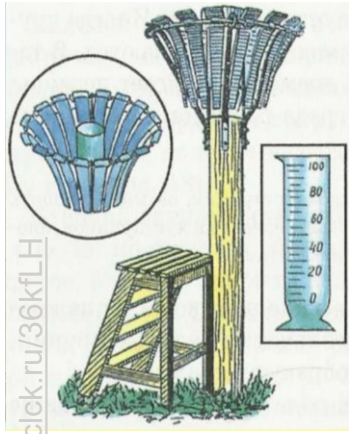
квантиль 0.95

Признаки (предикторы)

27 метеостанций
май – сентябрь
1989 – 2020 гг.



Осадкомер Третьякова



Модели

Методика и материалы:

35 физически обоснованных **предикторов** осадков

Постановка задачи

Целевая переменная

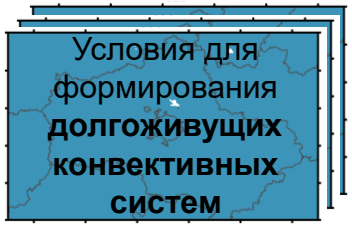
Признаки (предикторы)

Модели

Данные реанализа ERA5
(осреднённые в домене за сутки)

- Интегральное влагосодержание тропосферы
- Абсолютная влажность в пограничном слое
- Абсолютная влажность в средней тропосфере
- Интегральная дивергенция водяного пара

- Доступная потенциальная энергия конвекции
- Наибольшая разность потенциальных температур на поверхности и в тропосфере
- Температура в слоях тропосферы
- Вертикальный градиент температуры в средней тропосфере
- Фронтальный параметр
- Модуль градиента и лапласиана температуры на уровне 850 гПа



- Сдвиги ветра в слоях от поверхности до 1, 3 и 5 км.

- Модуль градиента и лапласиана приземного давления
- Давление на уровне моря
- Скорость и направление ветра у земли и в средней тропосфере
- Вертикальная скорость, дивергенция, геопотенциал в различных слоях тропосферы

Методика и материалы:

35 физически обоснованных **предикторов** осадков

Постановка задачи

Целевая переменная

Признаки (предикторы)

Модели

Данные реанализа ERA5
(осреднённые в домене за сутки)

Abbr.	Name	Height	Unit
Moisture conditions			
PW	Total column water vapor content	Troposphere	kg m^{-2}
MLMR05	Specific humidity in boundary layer	Sfc – 500 m	kg kg^{-1}
DIVWV	Water vapor integral divergence	Troposphere	kg m^{-2} s^{-2}
q_500	Specific humidity	500 hPa	kg kg^{-1}
tp_mean	Total precipitation	sfc	mm
Thermal conditions			
CAPE (SB, ML, MU)	Convective available potential energy	Troposphere	J kg^{-1}
d_thetaPE	Differences in pseudoequivalent temp.	Sfc – height of min(Θ_e)	K
t	Temperature	850, 700, 500 hPa	K
LR58	lapse rate	500 – 850 hPa	K km^{-1}
Dynamic conditions for deep convection occurrence			
LLS	low level shear	sfc – 1 km	m s^{-1}
MLS	mid level shear	sfc – 3 km	m s^{-1}
DLS	deep layer shear	sfc – 5 km	m s^{-1}
Circulation-related predictors			
TFP	Thermal front parameter	Sfc – 300 hPa	-
DT/GT	Laplacian and temperature gradient modulus	850 hPa	K..
DP/GP	Laplacian and pressure gradient modulus	Sfc	hPa..
msl	Mean sea level pressure	m.s.l.	hPa
Wind	Wind speed & direction	10 m., 500 hPa	m s^{-1}
w	Vertical velocity	850, 700, 500 hPa	Pa s^{-1}
d	Divergence	950, 500, 300 hPa	s^{-1}
z	Geopotential	850, 700, 500 hPa	m^2 s^{-2}

Методика и материалы:

Постановка задачи

Целевая переменная

Признаки (предикторы)

Модели

Характеристики влажности

Интегральное* влагосодержание
– влагозапас атмосферы, потенциал формирования осадков

$$PW = \frac{1}{\rho g}\int_{p_1}^{p_2} q(p)dp,$$

Средняя влажность в погранслое (0 – 500 м)
– косвенная характеристика интенсивности начальной конденсации при конвекции

Интегральная* дивергенция водяного пара
– показывает, действуют ли потоки в атмосфере на увеличение влагозапаса.

* В толще тропосферы

Методика и материалы:

Постановка задачи

Термическая стратификация, термодинамическое состояние атмосферы

Целевая переменная

Доступная потенциальная энергия конвекции

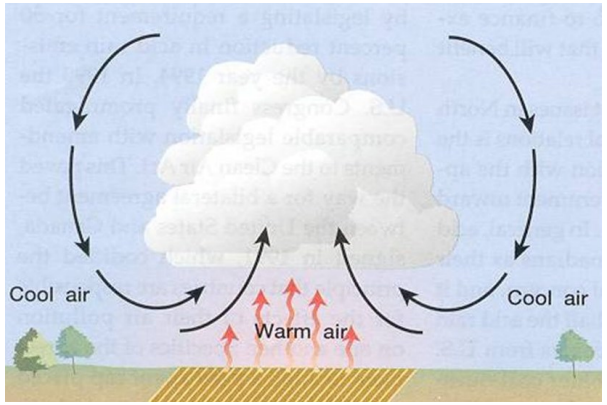
– мера способности атмосферы поддерживать восходящие движения воздуха

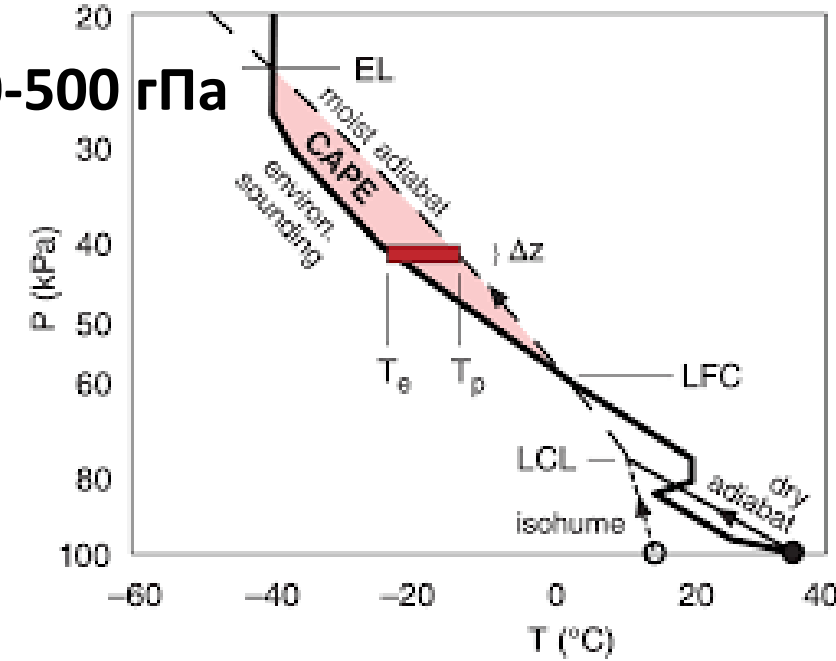
$$CAPE = g \int_{p(LFC)}^{p(EL)} \frac{T_{v,p} - T_{v,e}}{\bar{T}_{v,e}} dp$$

Признаки (предикторы)

Градиент температуры в слое 850-500 гПа

– неустойчивость в средней тропосфере





Модели

Методика и материалы:

Постановка задачи

Целевая переменная

Признаки (предикторы)

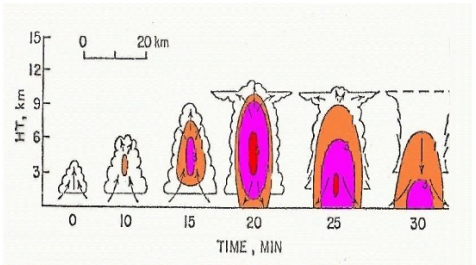
Модели

Условия для организации конвекции

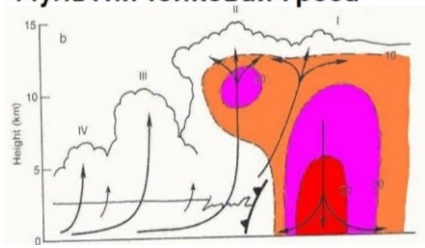
Сдвиги ветра в слоях от поверхности до 1, 3, 6 км

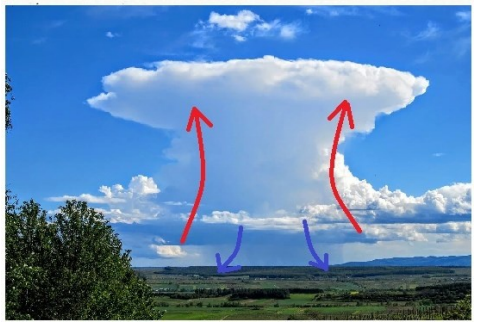
– условия для формирования мультиячейковых и суперъячейковых систем (пространственная дифференциация восходящих и нисходящих потоков в облаке, увеличивающая продолжительность жизни облака)

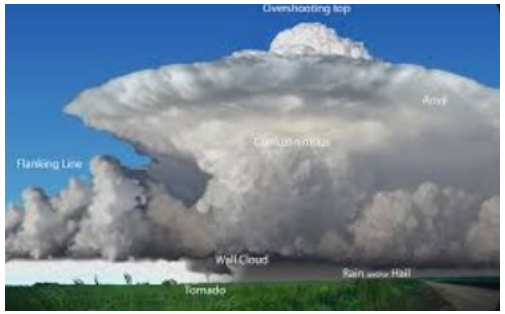
Моноячейковая гроза

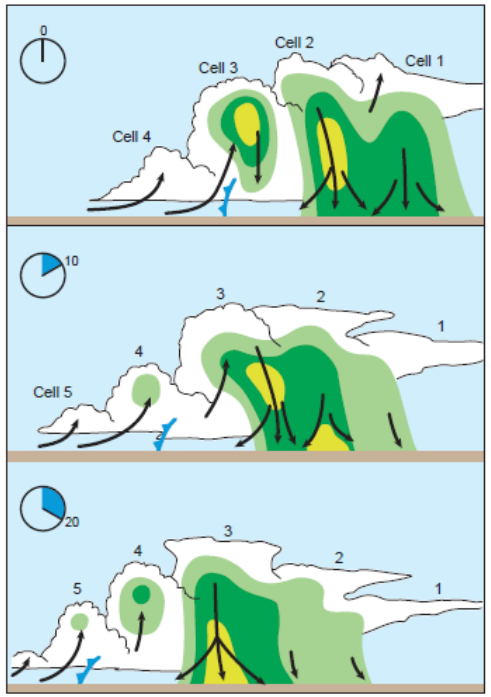


Мультиячейковая гроза









Методика и материалы:

Постановка задачи

Целевая переменная

Признаки (предикторы)

Модели

Циркуляционные предикторы

Лапласиан приземного давления

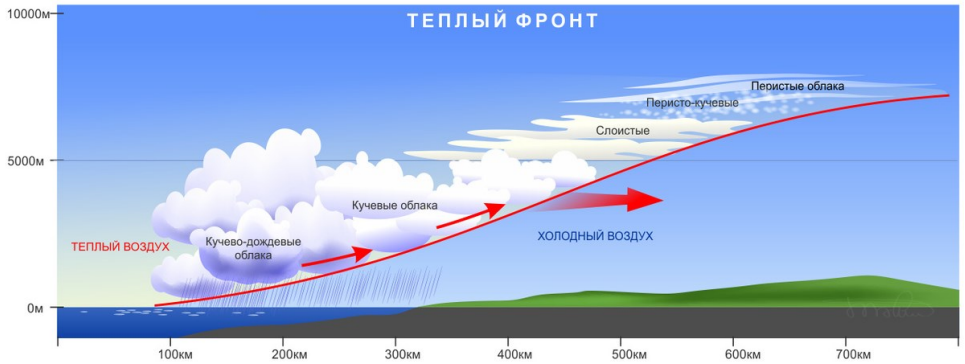
– характеризует циклоничность (крупномасштабную конвергенцию) воздушных масс в приземном слое

Лапласиан/градиент температуры на высоте 850 гПа (~3 км)

– характеризует степень бароклинности в средней тропосфере → наличие/отсутствие фронтальных зон

Фронтальный параметр

– характеристика бароклинности и непосредственного наличия и интенсивности фронта



Методика и материалы:

Постановка задачи

Целевая переменная

Признаки (предикторы)

Модели

Гребневая регрессия

(L_2 – регуляризация, регуляризация Тихонова)

$$\widehat{\beta^{ridge}} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}'$$

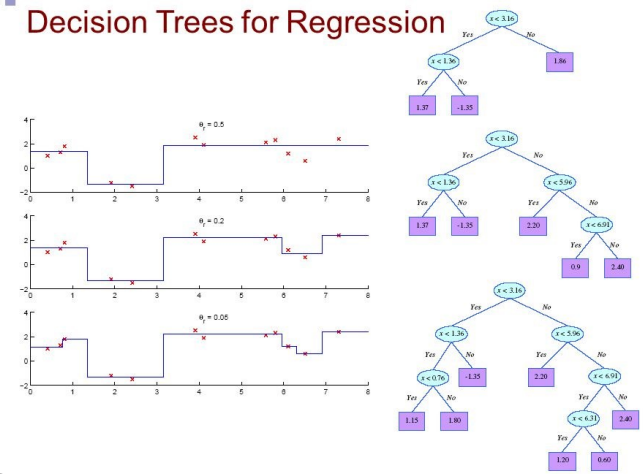
многомерная регрессия — штраф

N - количество наблюдений
p – количество предикторов
y_i – *i*-ое наблюдение целевой переменной
x_{ij} – *i*-ое наблюдение *j*-го предиктора
β_0 – свободный член
β_j – коэффициенты модели
λ – параметр регуляризации

Случайный лес

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x)$$

Решающие деревья равноправны



N – количество деревьев;
i – счетчик для деревьев;
b – решающее дерево;
x – сгенерированная нами на основе данных выборка.

Градиентный бустинг

Решающие деревья неравноправны: последний оценщик, которому передаётся информация предыдущих, имеет больший вес.

Результат:

наиболее оптимальная конфигурация эксперимента

Целевая переменная

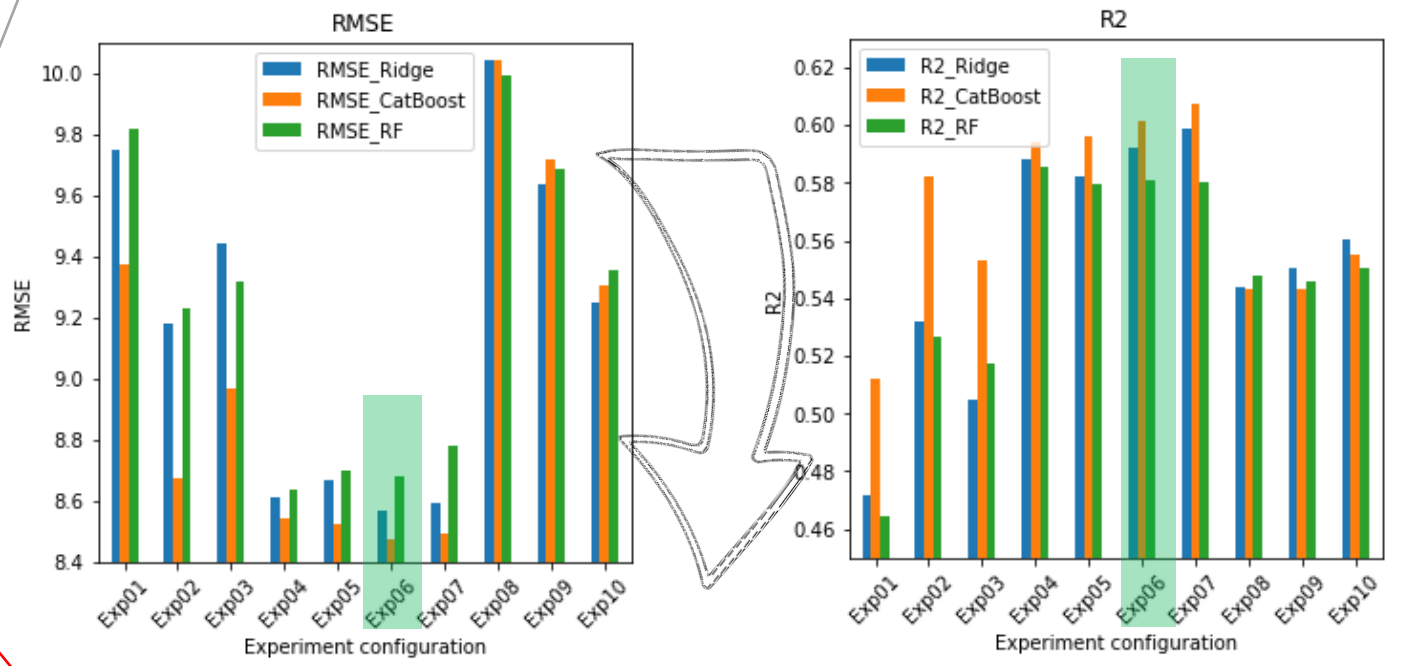
Максимальная в регионе суточная сумма осадков

Table 4. Experiments' configuration

Sample name	Temporal averaging	Model tp_mean	Normalization	Train-test split	Cross-validation	
					Iterations N	Shuffling
Exp01	*daily max*	false	-	80/20	1	false
Exp02	*daily max + mean*	false	-	80/20	1	false
Exp03	*daily mean*	false	-	80/20	1	false
Exp04	daily mean	*true*	-	80/20	1	false
Exp05	daily mean	true	*minmax*	80/20	1	false
Exp06	daily mean	true	*mean*	80/20	1	false
Exp07	daily mean	true	mean	*75/25*	1	false
Exp08	daily mean	true	mean	80/20	1	*true*
Exp09	daily mean	true	mean	80/20	10	true
Exp10	daily mean	true	mean	80/20	20	true

Начало проведения экспериментов:

признаки, осреднённые за сутки; без признака осадков; без нормализации



Результат (опорный метод):

признаки, осреднённые за сутки; с признаком осадков; с нормализацией

Результат:

наиболее оптимальная конфигурация эксперимента

Целевая переменная

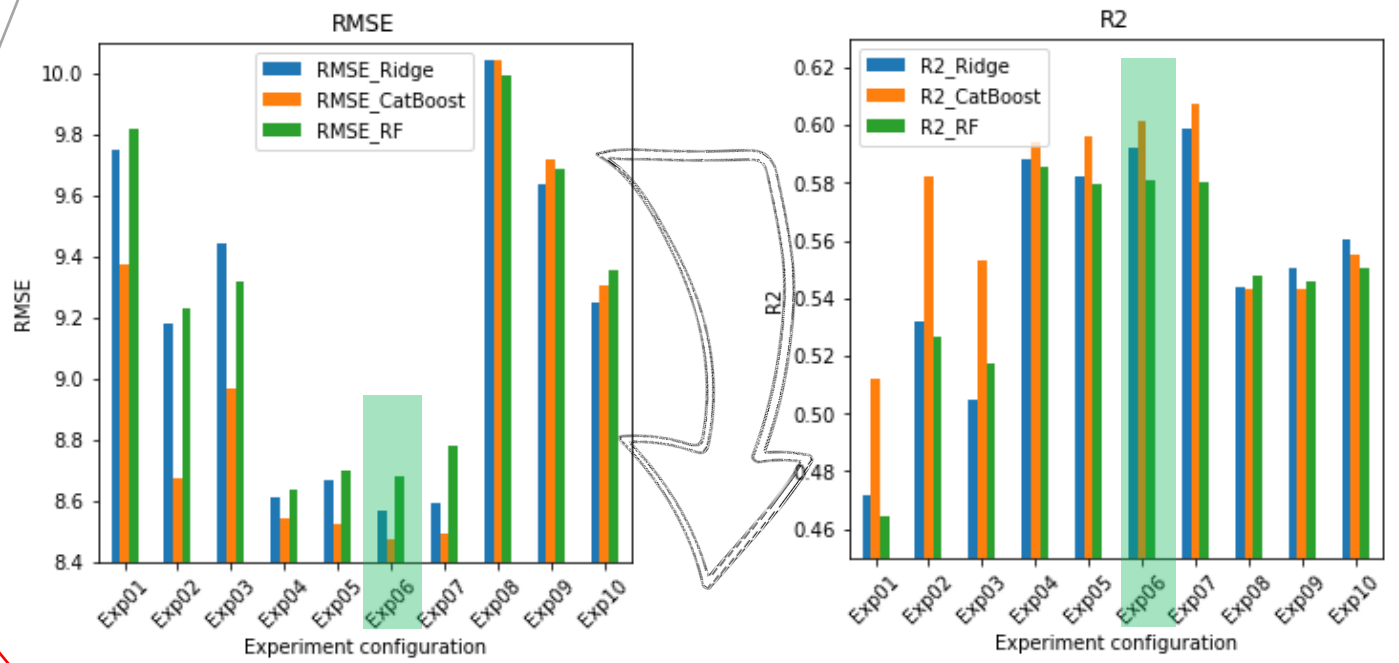
Максимальная в регионе суточная сумма осадков

Table 4. Experiments' configuration

Sample name	Temporal averaging	Model tp_mean	Normalization	Train-test split	Cross-validation	
					Iterations N	Shuffling
Exp01	*daily max*	false	-	80/20	1	false
Exp02	*daily max + mean*	false	-	80/20	1	false
Exp03	*daily mean*	false	-	80/20	1	false
Exp04	daily mean	*true*	-	80/20	1	false
Exp05	daily mean	true	*minmax*	80/20	1	false
Exp06	daily mean	true	*mean*	80/20	1	false
Exp07	daily mean	true	mean	*75/25*	1	false
Exp08	daily mean	true	mean	80/20	1	*true*
Exp09	daily mean	true	mean	80/20	10	true
Exp10	daily mean	true	mean	80/20	20	true

Начало проведения экспериментов:

признаки, осреднённые за сутки; без признака осадков; без нормализации



Результат (опорный метод):

признаки, осреднённые за сутки; с признаком осадков; с нормализацией

ERA5 0.25 x 0.25

RMSE = 9.8

R^2 = 0.46

Exp06 (ML)

RMSE = 8.5 – 8.6

R^2 = 0.58 – 0.6

Качество статистической детализации

Результат:

реализация опорного метода расчёта характеристик пространственного распределения осадков

Значения показателей эффективности обученных моделей статистической детализации на тренировочных выборках (2015 – 2020 гг.)

Target variables	Range of parameters, mm	ML MODEL & quality metrics					
		Ridge		CatBoost		Random Forest	
		RMSE	R2	RMSE	R2	RMSE	R2
Max	117.7	**8.59**	**0.59**	**8.52**	**0.6**	**8.71**	**0.58**
Std	30.4	2.13	0.59	2.09	0.61	2.14	0.59
Kurt	23.7	5.79	0.35	5.81	0.34	5.77	0.35
Skew	6.2	1.03	0.45	1.03	0.45	1.01	0.47
Q 0.9	89	5.26	0.65	5.35	0.64	5.35	0.64
Q 0.95	99.2	6.8	0.61	6.87	0.61	6.89	0.6

Результат:

Значимость признаков для целевой переменной максимальной суточной суммы осадков

Целевая переменная

Максимальная в регионе суточная сумма осадков

SHAP (SHapley Additive exPlanations)

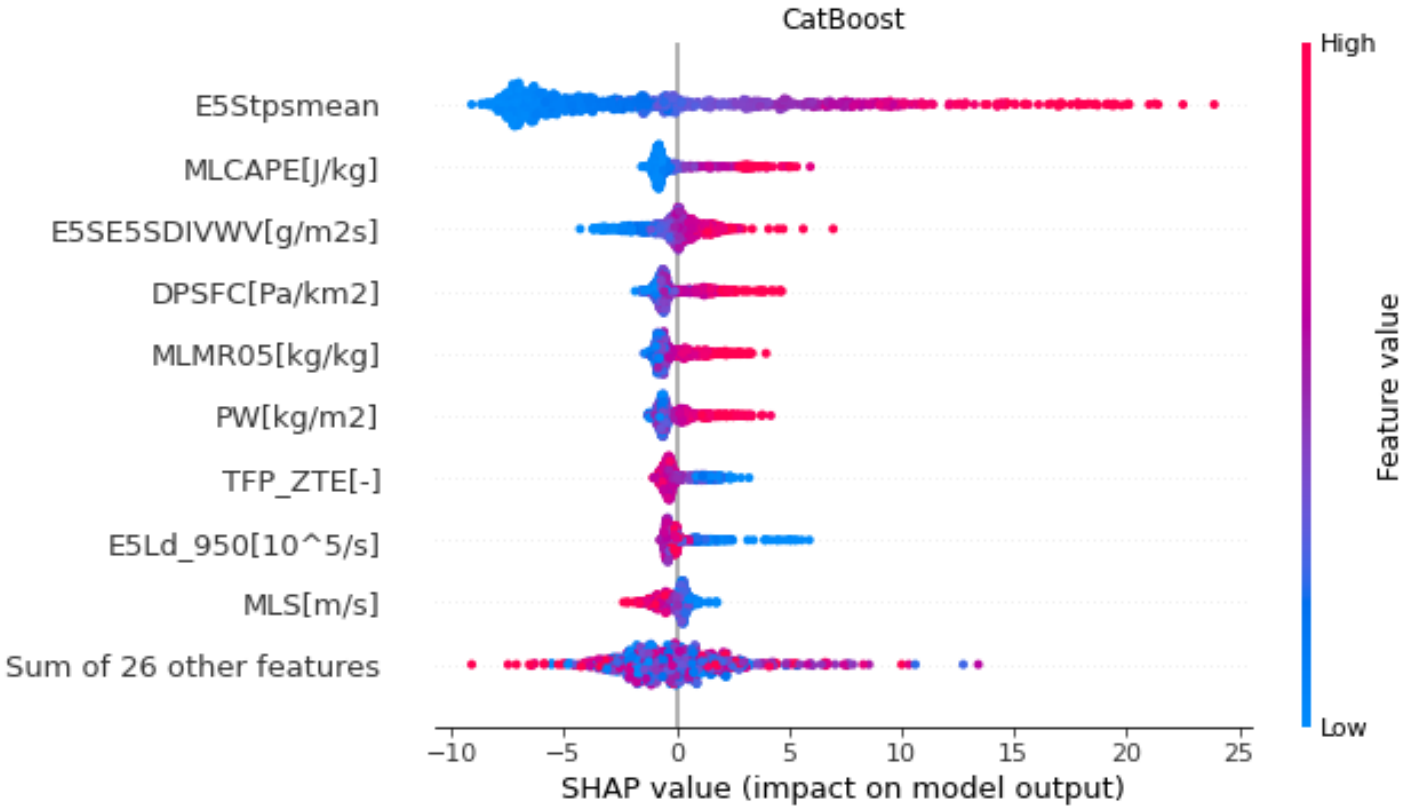
$$\phi_i(p) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!}(p(S \cup \{i\}) - p(S))$$

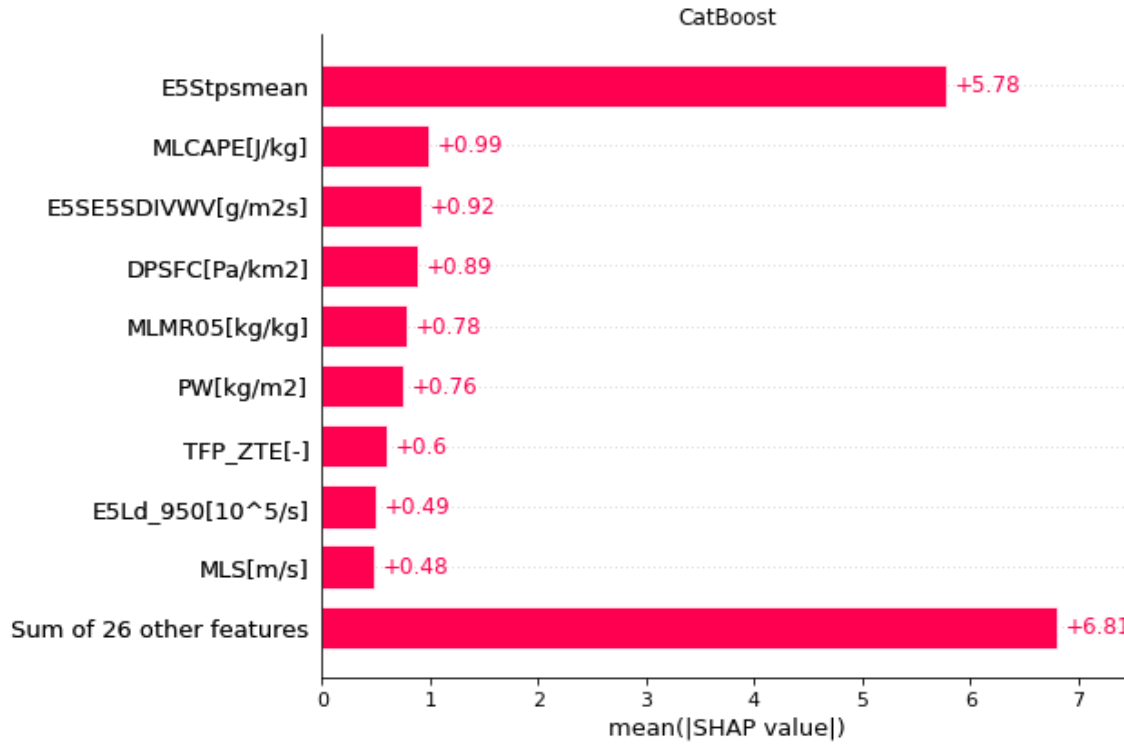
$p(S \cup \{i\})$ — это предсказание модели с i-той фичей,

$p(S)$ — это предсказание модели без i-той фичи,

n — количество фичей,

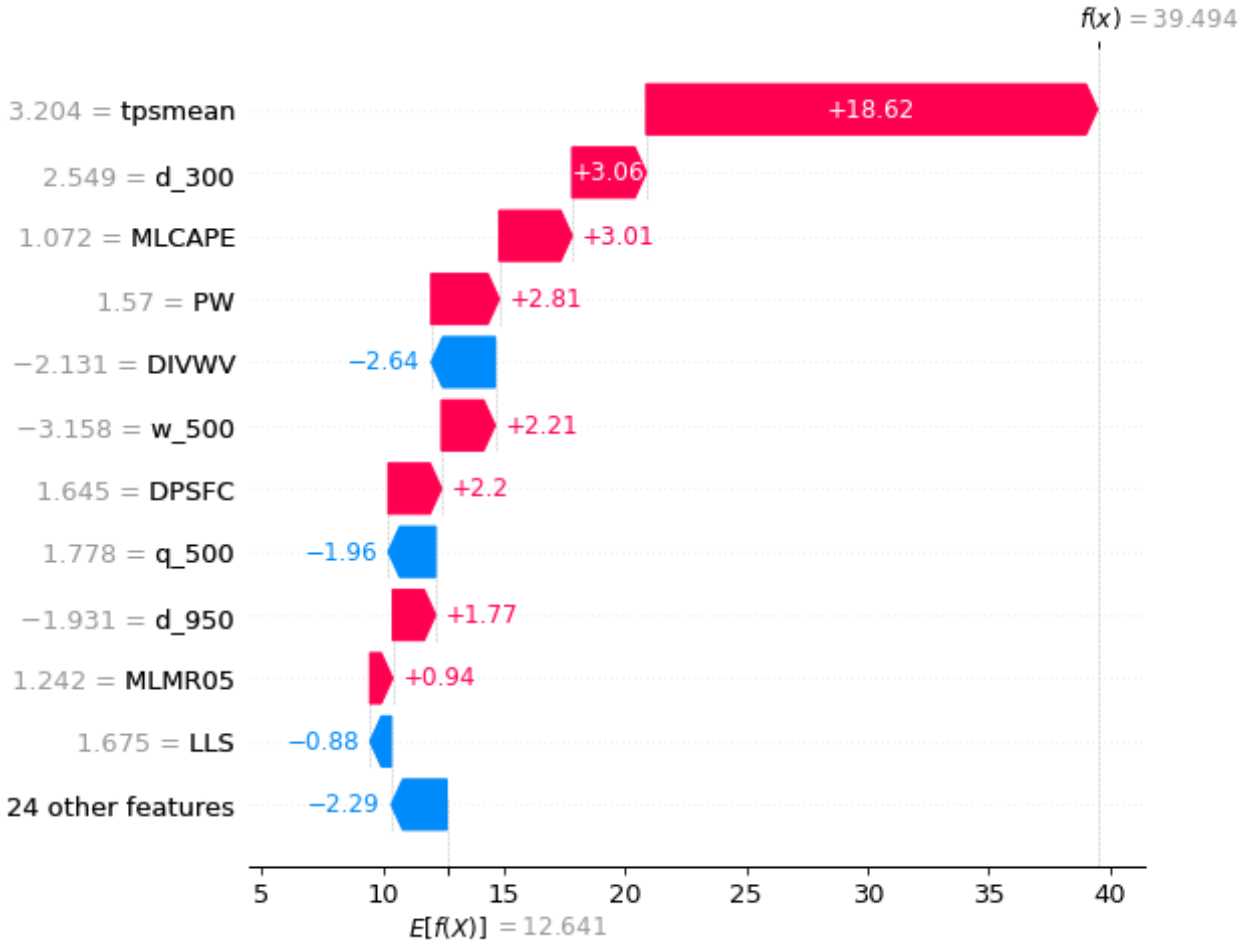
S — произвольный набор фичей без i-той фичи

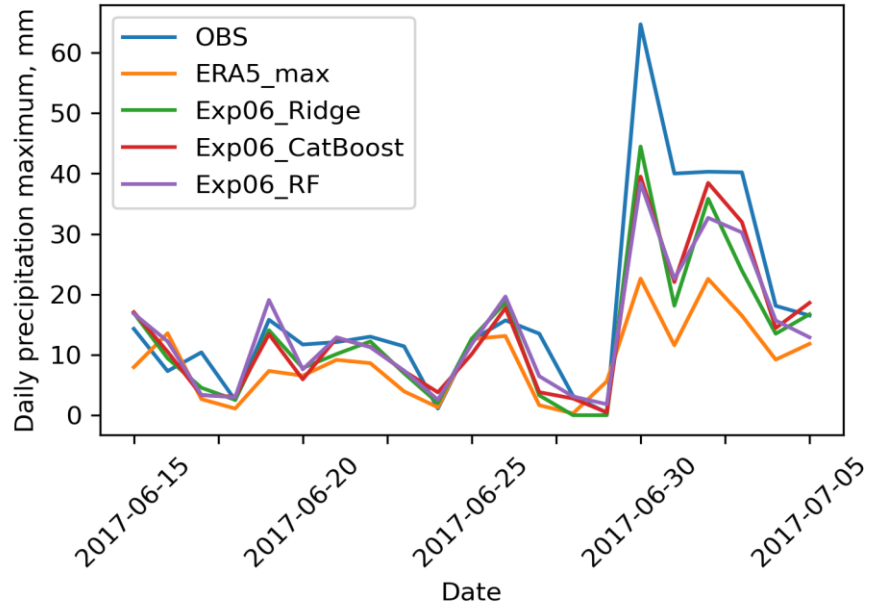


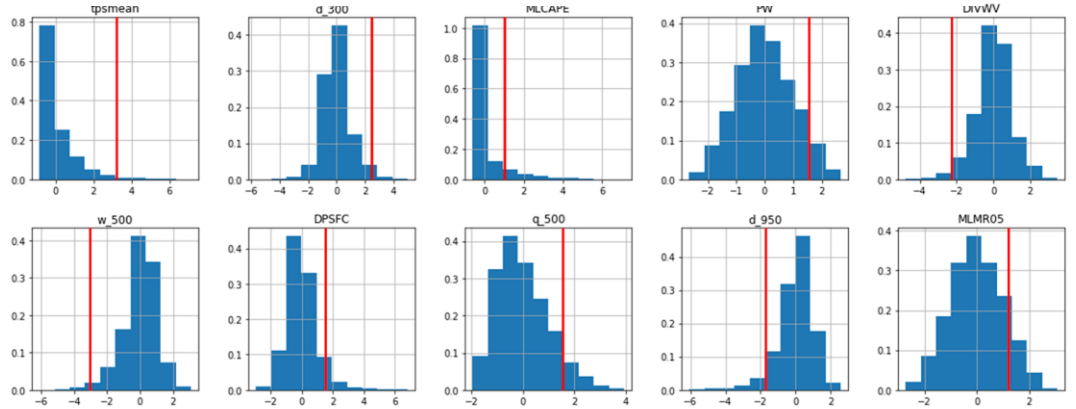


Результат:

применение для анализа случая сильных осадков 30.06.2017







Выводы:

- После проведённых экспериментов был выбран опорный метод и проведены эксперименты для характеристик пространственного распределения осадков.
- Среднеквадратическая ошибка опорного метода (максимальная в регионе суточная сумма осадков) составила 8.59 мм для модели гребневой регрессии и 8.52 мм для модели градиентного бустинга, 8.71 для модели случайного леса (менее 10% от общего разброса величин).
- Рейтинг значимости признаков заметно варьировался в зависимости как от конфигурации эксперимента (выбора среднего/максимума признаков за сутки, учёте/неучёте модельного признака осадков, типа нормализации), так и от выбранной модели машинного обучения. В целом в топ-10 признаков входят все группы признаков по характеру воздействия на системы осадков: термодинамические характеристики, параметры неустойчивости, влагосодержания и конвергенции.

В перспективе:

- Оценка применимости для других регионов
- В качестве целевой переменной – параметры теоретического пространственного распределения осадков (центральные моменты)
- Получение оптимального набора предикторов на основе ансамбля экспериментов

Спасибо за внимание!