Машинное обучение в гидрометеорологии

Лекция №1. Введение. Организационные вопросы.

Михаил Иванович Варенцов (mikhail.varentsov@srcc.msu.ru)

Михаил Алексеевич Криницкий (krinitsky@sail.msk.ru)

ml4hydromet@ml4es.ru

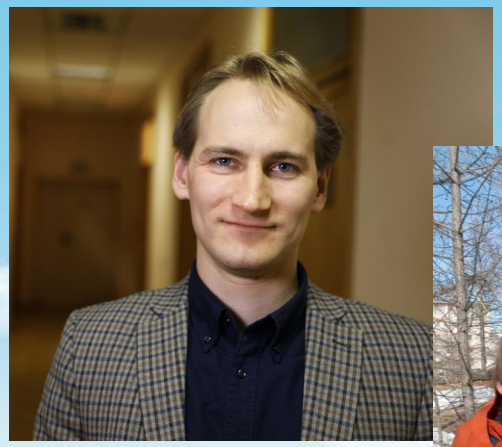
ИНТЕЛЛЕКТ

ФОНД РАЗВИТИЯ НАУКИ И ОБРАЗОВАНИЯ

Давайте познакомимся

Коротко обо мне:

- Выпускник географического факультета МГУ, кафедры метеорологии и климатологии (2014)
- Кандидат географических наук (2018), тема диссертации «Анализ и моделирование мезоклиматических особенностей Московской агломерации» (научный руководитель – проф., д.г.н. А.В. Кислов)
- Старший научный сотрудник НИВЦ МГУ, а также сотрудник ИФА РАН, Гидрометцентра РФ и т.д.
- Автор более 70 публикаций в международных рецензируемых журналах
- **Компетенции и интересы:**
  - Городская метеорология и климатология
  - Численное моделирование погоды и климата на региональном масштабе
  - Работа на суперкомпьютерах
  - Анализ данных
  - Экспериментальные метеорологические исследования



Информация о курсе

- **Курс экспериментальный, на географическом факультете впервые**
- **Разработан на базе курса «Машинное обучение в науках о Земле» Михаила Криницкого (https://github.com/MKrinitskiy/ML4ES*)**
- **12 занятий по 2 пары (лекция + семинар)**
- **Формы контроля успеваемости:**
  - Практические работы, они же домашние задания (9 шт.)
  - Курсовой проект – обобщение результатов практических работ.
    В конце курса курса будет защита проектов.
  - Экзамен (с возможность получения «автомата»)

Коммуникация

- Почта для отправки практических работ: **ml4hydromet@ml4es.ru**
- Группа курса ТГ: https://t.me/+n5MY6B6olJM4OWYy
- Репозиторий с материалами по курсу: https://github.com/mvarentsov/ML4hydromet-2024

План курса

1. Введение (сегодня)
2. Технические средства анализа данных и машинного обучения
3. Вероятностная постановка задач МО
4. Методология подхода машинного обучения
5. Обзор основных методов машинного обучения
6. Краткое знакомство с методами глубокого обучения
7. Применение методов МО в гидрометерологии
8. Методы МО для решения задач типа "Обучение без учителя"

Практические работы

- **Практические работы (домашние задания) в основном заключаются в применении разбираемых в рамках курса методов и моделей к выбранным наборам данных.**
  Примеры заданий:
  - Визуализация данных и их распределений
  - Применение линейной или логистической регрессии
  - Применение полносвязной искусственной нейронной сети

Курсовой проект – обобщение совокупности практических работ

- **Требования к выполнению практических работ:**
  - Выполнение на языке программирования Python
  - Оформление кода и результатов в формате Jupiter Notebook
- **Оценка за каждую практическую работу определяется:**
  - Исходным код, использованным для анализа данных
  - Визуализацией результатов
  - Текстовым описанием методов и результатов
  - Ответами на вопросы по теме практической работы
- **Мягкие дедлайны, жесткие штрафы за сдачу после дедлайна ☺**

python™

jupyter

Готовые коллекции данных

Название и ссылка	Предметная область	Куратор	Описание возможных задач
DASIO (All-Sky Imagery over the Ocean)	Атмосферная радиация и облачность	Михаил Криницкий	Аппроксимация радиационных потоков на основе фотоснимков небесной полусферы
DISO3 (In-Situ Observations over the Ocean)	Приводный слой атмосферы	Михаил Криницкий	Анализ взаимосвязей между метеовеличинами
DDM (Discharge and Meteorology)	Гидрология	Михаил Варенцов	Аппроксимация и прогноз расхода воды в реке на основе осредненных по водосбору характеристиках метеорологического режима по данным реанализа
DUHI (Urban Heat Island)	Городская метеорология	Михаил Варенцов	Аппроксимация и прогноз интенсивости городского острова тепла - разности температуры между городом и фоновым ландшафтом
DCIPP (Dataset of Convective Intensive Precipitation Predictors)]	Метеорология	Юлия Ярынич	Оценка рисков возникновения опасных явлений, связанных с глубокой конвекцией

Атмосферная радиация и облачность

Dataset of **A**ll-**S**ky **I**magery over the **O**cean

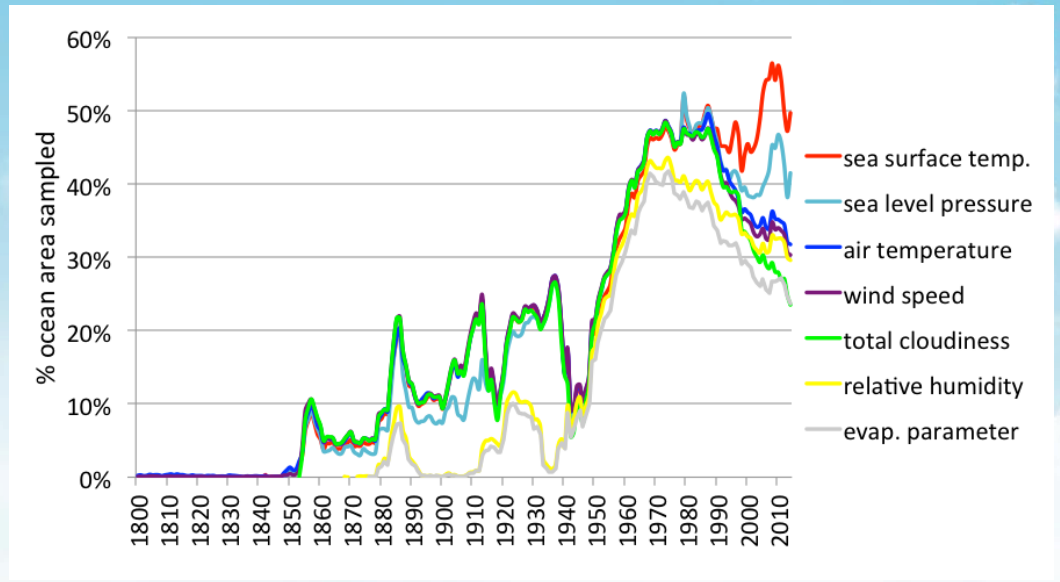
- `photo_name` - имя файла фотографии
- `photo_datetime` - дата и время регистрации фотографии (UTC)
- `CM3up[W/m2]` - поток полной приходящей коротковолновой радиации (Вт/м2)
- `CG3up[W/m2]` - поток полной приходящей длинноволновой радиации (Вт/м2)
- `CM3down[W/m2]` - поток полной уходящей коротковолновой радиации (Вт/м2)
- `CG3down[W/m2]` - поток полной уходящей длинноволновой радиации (Вт/м2)
- `radiation_datetime` - дата и время момента регистрации осредненных за 10с радиационных потоков
- `feature0`, `feature1`, ... `feature26` - статистические признаки красного (R) цветового канала фотографии за исключением незначимых областей снимка (надстроек парохода, угловых зон снимка). Рассчитываются следующие признаки:
  - `min` (минимальное значение)
  - `max` (максимальное значение)
  - `mean` (выборочное среднее)
  - `var` (выборочная дисперсия)
  - `skew` (выборочная оценка коэффициента асимметрии)
  - `kurt` (выборочная оценка коэффициента эксцесса)
  - `p1`, `p5`, `p10`, `p15`, ... `p95`, `p99` (итого 21 шт) - выборочные оценки эмпирических перцентилей уровней 1%, 5%, 10% далее с шагом 5% до 95%, а также перцентиля уровня 99%.
- `feature27` - `feature53` - статистические признаки зеленого (G) цветового канала фотографии

Krinitskiy, M.; Aleksandrova, M.; Verezemskaya, P.; Gulev, S.; Sinitsyn, A.; Kovaleva, N.; Gavrikov, A. *On the Generalization Ability of Data-Driven Models in the Problem of Total Cloud Cover Retrieval*. Remote Sens. **2021**, *13*, 326. https://doi.org/10.3390/rs13020326 [2] Krinitskiy, M.; Koshkina, V.; Borisov, M.; Anikin, N.; Gulev, S.; Artemeva, M. *Machine Learning Models for Approximating Downward Short-Wave Radiation Flux over the Ocean from All-Sky Optical Imagery Based on DASIO Dataset*. Remote Sens. 2023, 15, 1720. https://doi.org/10.3390/rs15071720

Метеорология приводного слоя атмосферы

Dataset of In-Situ Observations over the Ocean

No кол.	Рек. имя	Ширина в символах	Тип данных	описание переменной
1	year	4	str	Номер года наблюдения
2	mon	3	str	Номер месяца наблюдения
3	day	3	str	Номер даты наблюдения
4	hour	5	str	Часовая компонента времени наблюдения
5	lat	10	float 7.2	Широта наблюдения
6	lon	9	float 6.2	Долгота наблюдения
7	hsun	7	float 4.2	Высота солнца над горизонтом
8	slp	8	float 5.2	атмосферное давление в гПа
9	ta	7	float 4.2	температура атмосферы, °C
10	sst	7	float 4.2	температура поверхности океана, °C
11	td	7	float 4.2	температура точки росы, °C
12	rh	8	float 4.3	относительная влажность (расчетное значение), в долях единицы
13	icn	3	int	Балл общей облачности, в октах (0-8); значение 9 кодирует состояние небосвода, при котором облака не могут быть наблюдены: туман, снег, др. метеорологические явления)



ICOADS

Городская метеорология

Dataset of Urban Heat Island (DUHI)

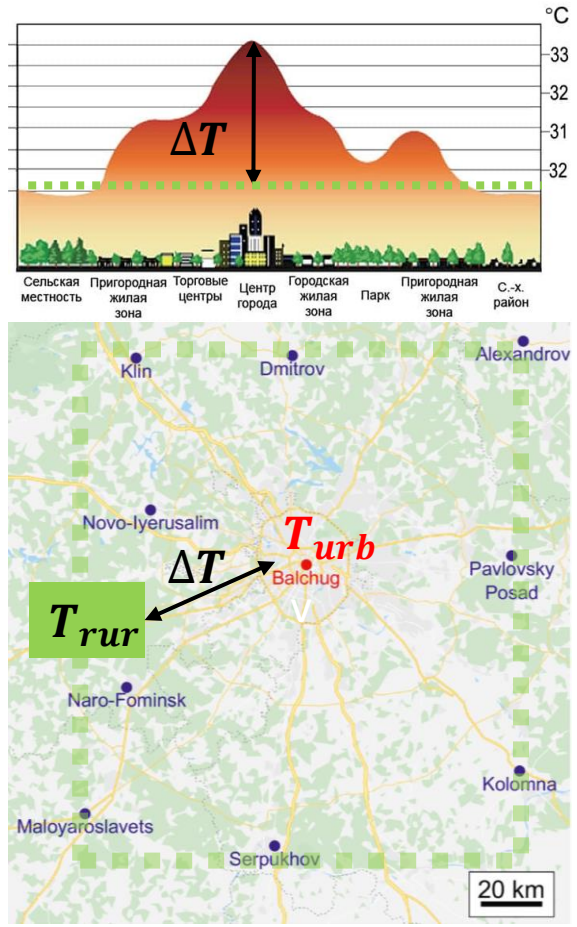
- **Целевая переменная:**
  разность температуры между центром Москвы (Балчуг) и средней температурой по 9 загородным метеостанциям
  $\Delta T = T_{urb} - T_{rur}$
- **Предикторы:**
  средние по региону характеристики метеорологического режима (реанализ ERA5 и/или наблюдения на загородных метеостанциях)

Varentsov, M.; Krinitskiy, M.; Stepanenko, V. *Machine Learning for Simulation of Urban Heat Island Dynamics Based on Large-Scale Meteorological Conditions*. Climate. **2023**, *11*, 200. https://doi.org/10.3390/cli11100200

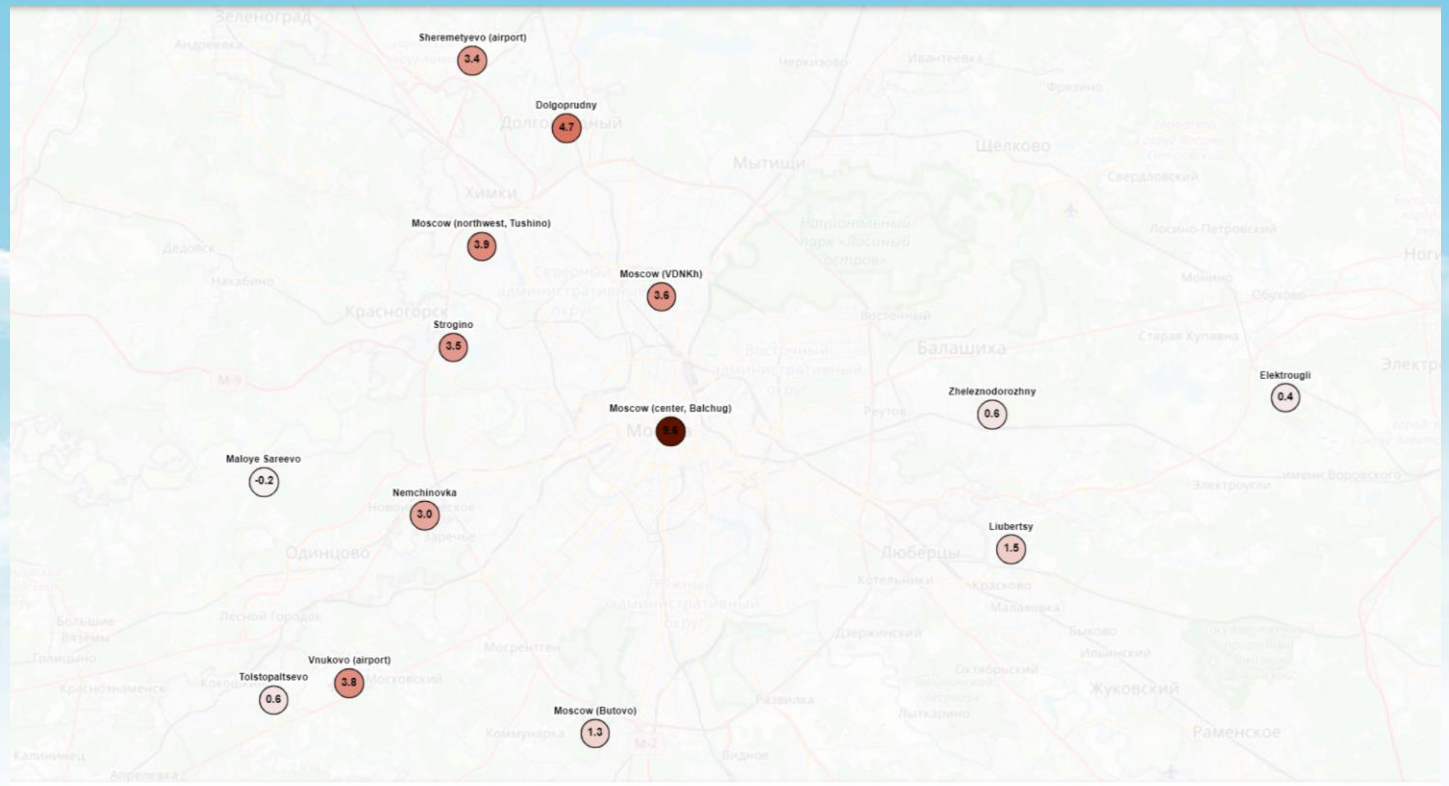
Низкоразрешенные метеоданные
(реанализ и/или наблюдения)

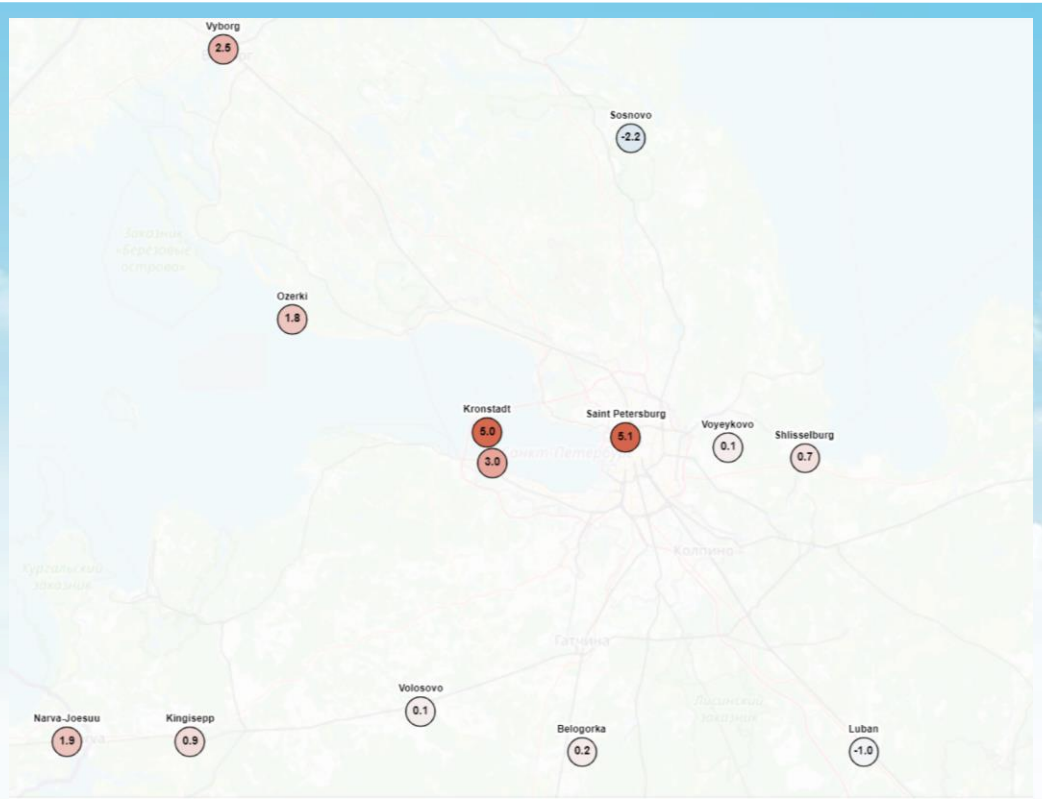
	Метеорологические предикторы	
Наблюдения и реанализ	**t2m**	Температура воздуха на 2 м, °C
	rh2m	Отн. влажность на 2 м, %
	vel10m	Скорость ветра на 10 м, м/с
	tcc	Доля общей облачности
	lcc	Доля нижней облачности
	WF	Фактор погоды, эмпирическая функция скорости ветра и облачности [Oke, 1998]
Только реанализ	**sp**	Атмосферное давление, гПа
	blh	Высота погранслоя, м
	str	Длинноволновый баланс, Вт/м^2
	ssr	Коротковолновый баланс, Вт/м^2
	strd	Приходящая коротковолновая радиация, Вт/м^2
	ssrd	Приходящая длинноволновая радиация, Вт/м^2
	tp	Сумма осадков за 3 ч, мм

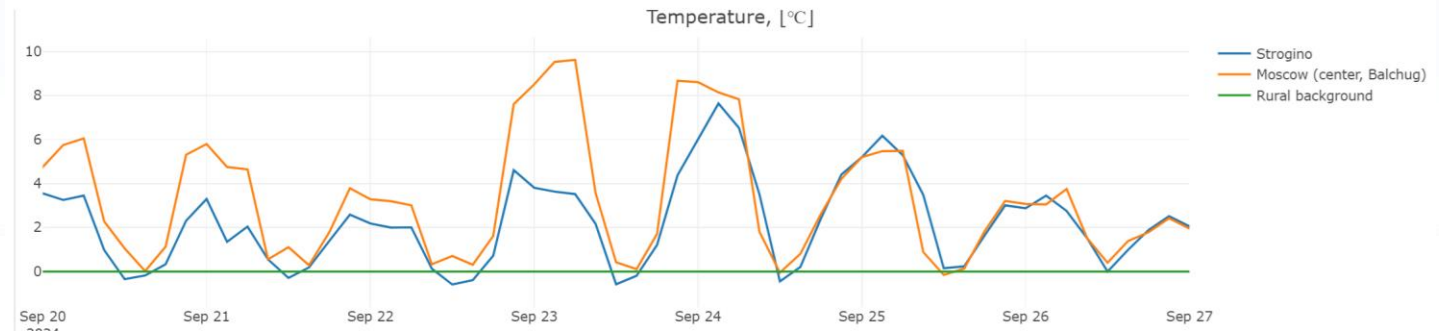
Наблюдаемая интенсивность острова тепла в точке

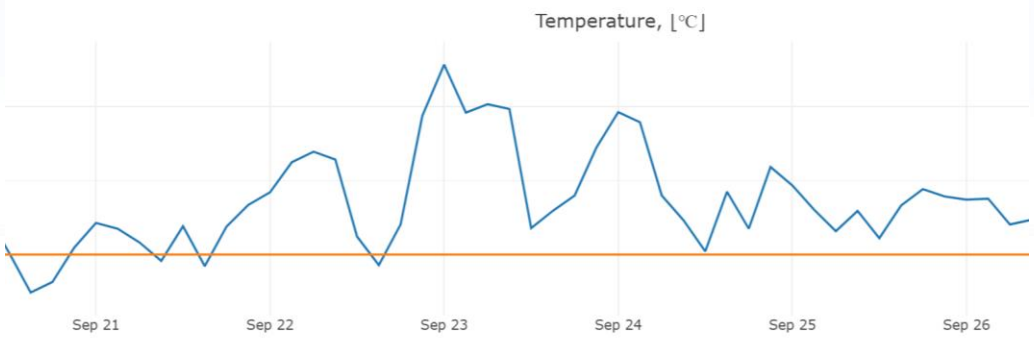


Городская метеорология





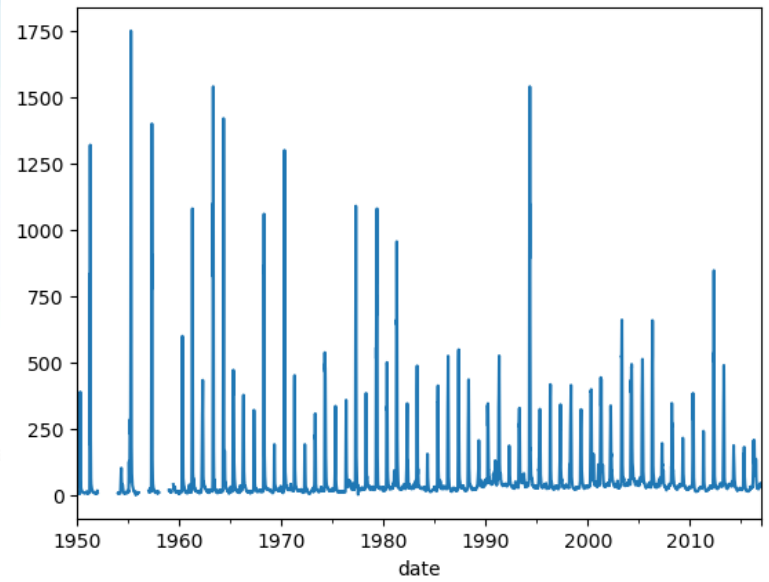


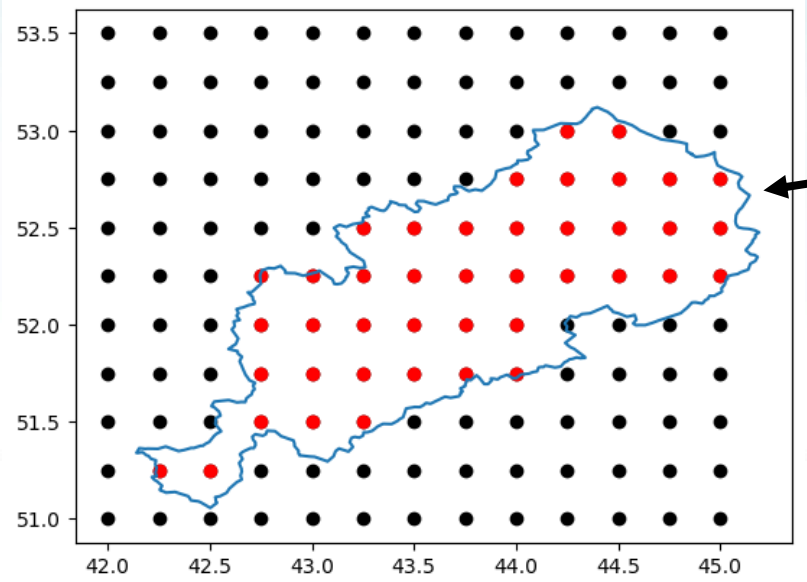


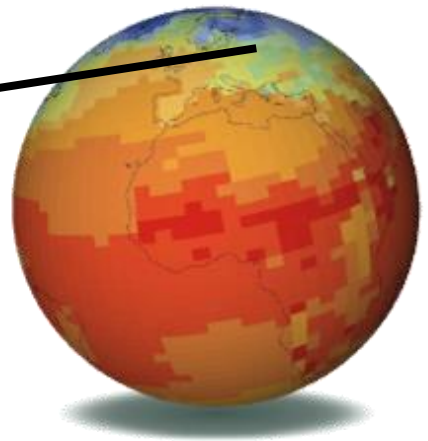
Гидрология

Dataset of **D**ischarge and **M**eteorology (DDM)

Река	Пост и ссылка на данные	Площадь водосбора (км2)	Временной период
Хопер	Пановка	932	1950-2016
Хопер	Балашов	14300	1950-2016
Хопер	Поворино	19100	1950-2016
Сосна	Елец	16300	1950-2016
Которосль	Гаврилов-Ям	4980	1980-2017







Конвективные осадки

Исходные данные реанализа

и др.

Расчёт ~50 производных характеристик состояния атмосферы

$$PW = \frac{1}{g}\int_{p1}^{p2} q(p)dp,$$

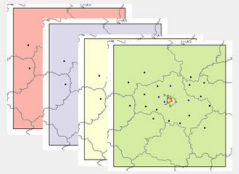
$$MLCAPE = g \int_{p(LFC-EL)} \frac{T_{v,p} - T_{v,e}}{\bar{T}_{v,e}} dp$$

$$TFP = \nabla|\nabla\, ZTE|\, \vec{n}_{ZTE}$$

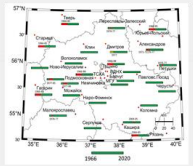
и др.

Итоговый архив признаков, осреднённых по площади

Набор Признаков из ERA5

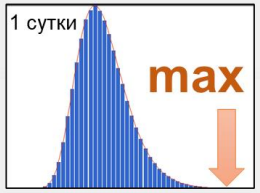


Данные метеостанций

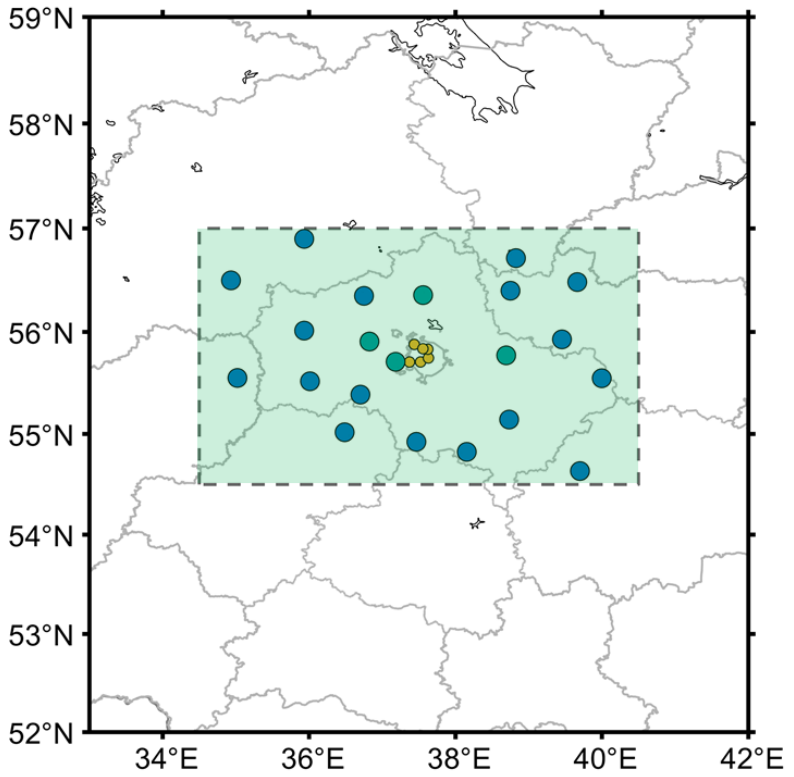


Модель машинного обучения

Максимальная суточная сумма осадков в регионе



Карта расположения метеостанций и домена ERA5



Метеостанции

- все станции фона (21)
- ближайшие станции фона (4)
- станции города (6)

область осреднения признаков

to be continued…

Домашняя работа №1

Нужно сформулировать задачу в терминах машинного обучения, которую можно было бы решать, имея в распоряжении доступные данные (одну из готовых коллекций данных или данные из вашей научной работы).

Для задачи следует указать следующие составляющие:

- класс задачи МО ("С учителем" / "Без учителя" / другое (уточнить))
- вид задачи МО (регрессия, кластеризация, понижение размерности, классификация, ...)
- целевая переменная, указать ее тип (категориальная, действительная, бинарная...) и размерность (количество значений на один объект/событие).
- описать функцию потерь, если это подразумевается типом задачи.
- описать объекты (события) в формулируемой задаче.
- предложить признаковое описание объектов/событий или описать уже имеющееся признаковое описание.
- предложить меру (меры) для оценки качества модели МО.

Не забываем, оформление результата – в формате markdown в Jupiter Notebook. Присылать на почту ml4hydromet@ml4es.ru

Итоговое описание задания будет в репозитории.

Давайте познакомимся

Расскажите о себе:

1) Как вас зовут?
2) Тема научного исследования (например, бакалаврского диплома)?
3) Идеи по применениию методов машинного обучения в вашем исследовании?