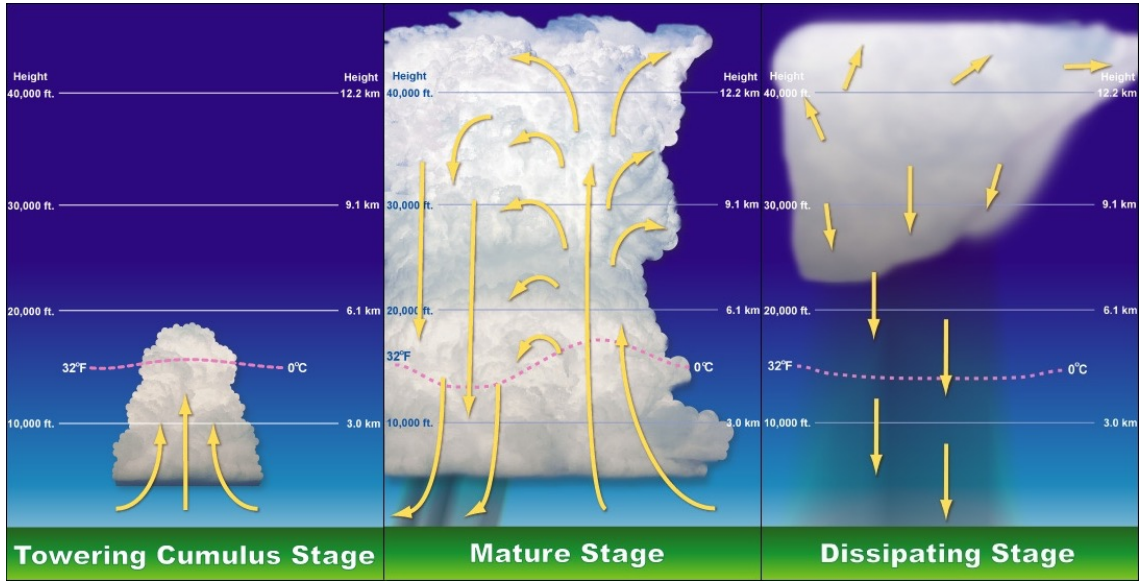
Обнаружение градоопасных состояний атмосферы с помощью машинного обучения на полях метеовеличин

П.Д. Блинов[1,2], А.В. Чернокульский[2], М.А. Криницкий[3,4], А.В. Бугримов[2]
[1]НИУ ВШЭ, [2]ИФА РАН, [3]МФТИ, [4]ИО РАН

Введение

- Град – одно из наиболее опасных природных явлений
- Традиционный пороговый подход – грубый, индексы не разрешаются климатическими прогнозами
- Статистика града в России – слабо изучена
- Изменение частоты градовых событий в будущем - не изучено



Ингредиенты конвективных явлений:

1) Теплый и влажный воздух внизу
2) Начальный вертикальный подъем
3) Неустойчивый профиль температуры
4) Сдвиг ветра

Индексы неустойчивости:

- CAPE (потенциальная энергия)
- LI (индекс плавучести)
- LLS, DLS (сдвиг ветра)
- ...

Цель

Разработать и сравнить интерпретируемые методы машинного обучения для задачи диагностики крупного града в России, на основе базовых метеовеличин

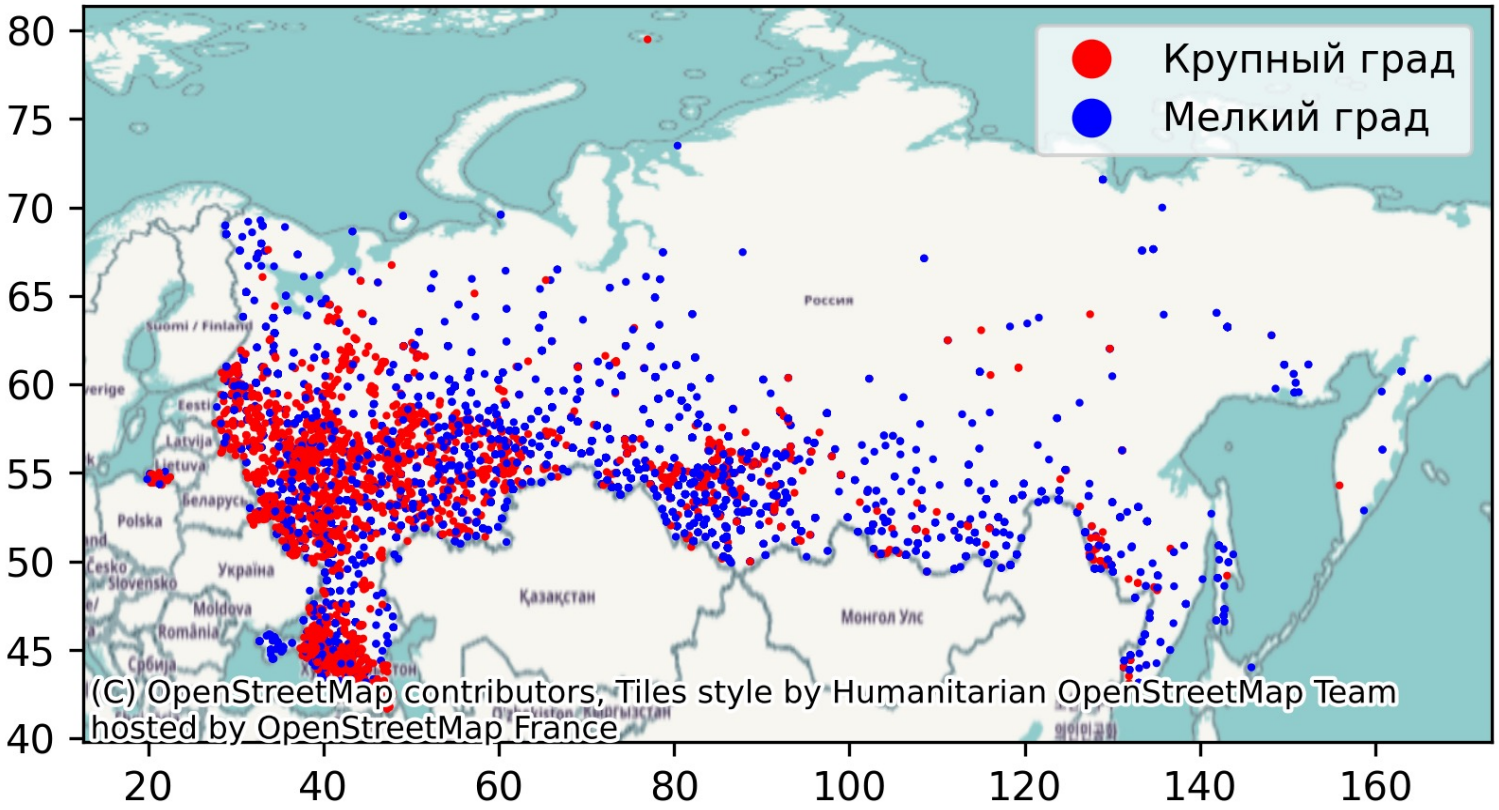
Задачи

1. Подготовка данных для обучения
2. Feature engineering
3. Оценка влияния параметров на качество работы моделей МО
4. Сравнение моделей между собой и с пороговым подходом
5. Сравнение методов на реальном случае крупного града

Новизна

- Большая часть работ занимается краткосрочным прогнозированием
- Обычно используют данные высокого разрешения
- Часто используют индексы
- Слабо раскрыт вопрос интерпретируемости
- На территории России исследований применимости ML в задаче не найдено

Исходные данные: случаи града



- База случаев града[1]:
  - 1996-2024
  - ESWD + ВНИИГМИ-МЦД
  - **7009** случаев мелкого града
  - **2354** случаев крупного града (размер от 2.0 см)
- Отрицательные случаи
  - Локации где наблюдения уже велись
  - Случайно выбранное время
  - Итого **34705** случаев без града

[1] (Бугримов и др., 2024)

Исходные данные: метеовеличины

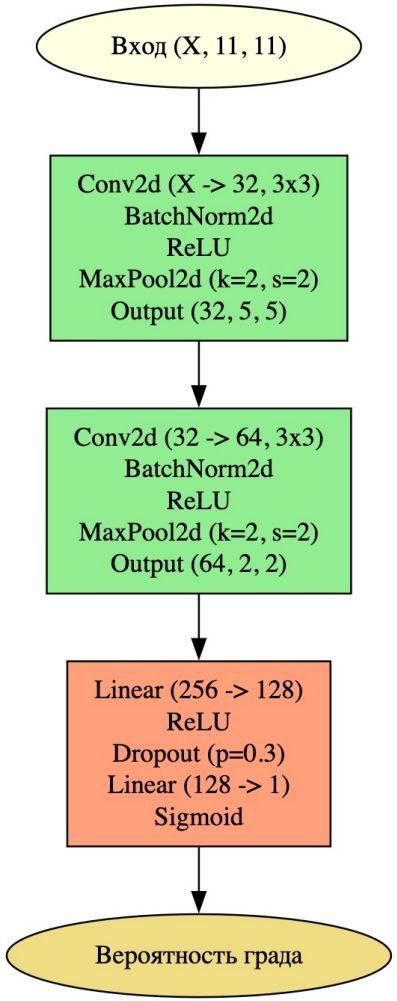
ERA5 (разрешение 0.25°, 1 ч):

- Геопотенциал
- Абсолютная влажность
- Температура
- Ветер (по компонентам)
- Относительная завихренность
- Температура точки росы (приземная)

Уровни: приземный + 850/700/500 гПа
Квадрат 8.25°x8.25° вокруг точки события (33x33 пикселя)

Модели: конфигурация

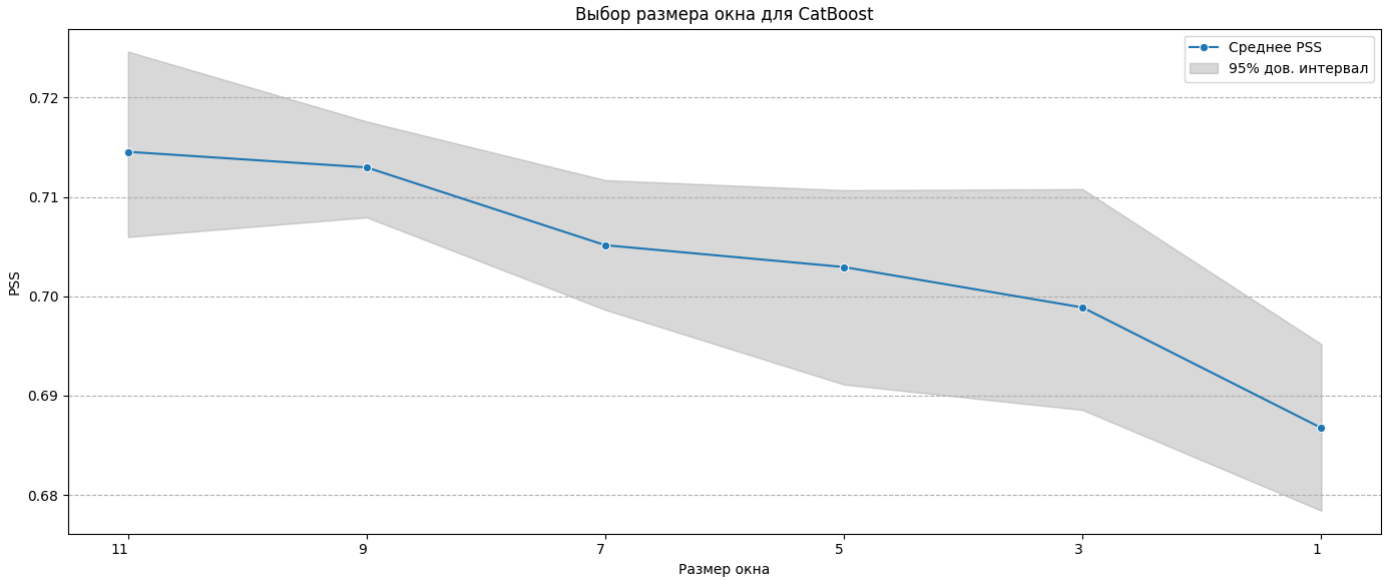
CNN



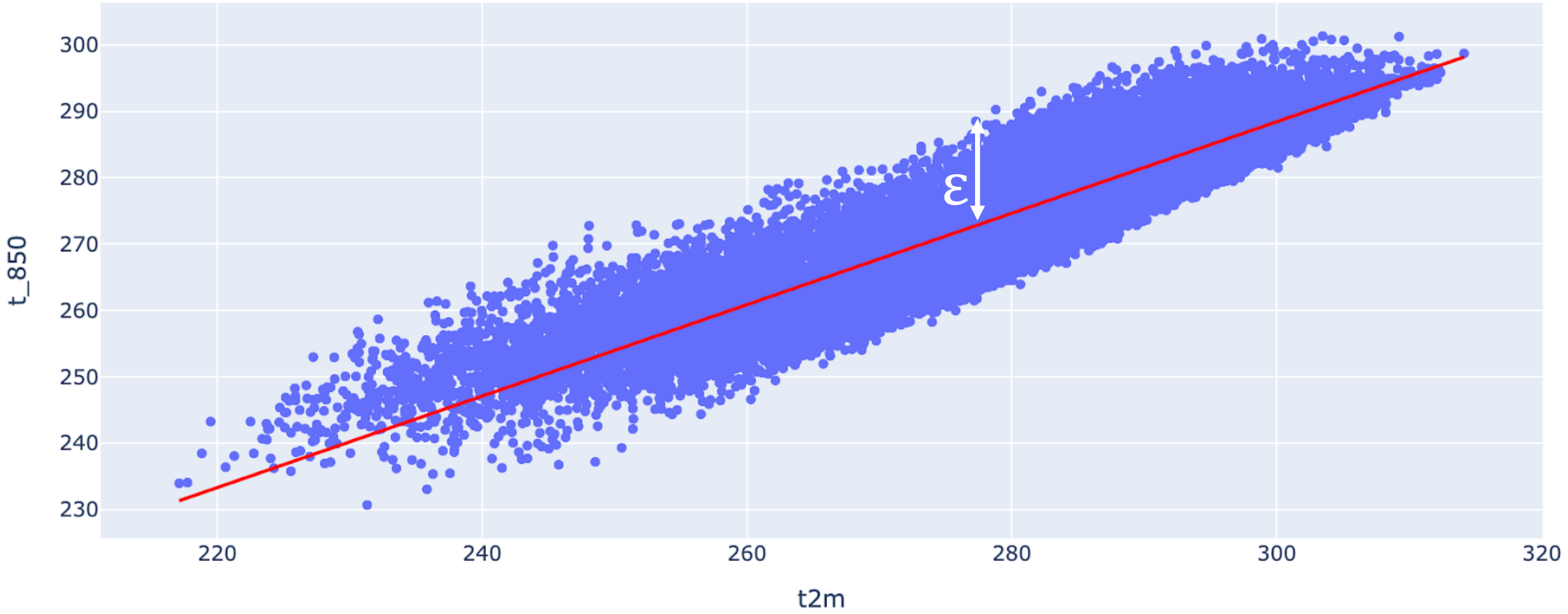
CatBoost

Из каждого входного слоя

- Минимум
- Максимум
- Среднее
- Значение в точке наблюдения

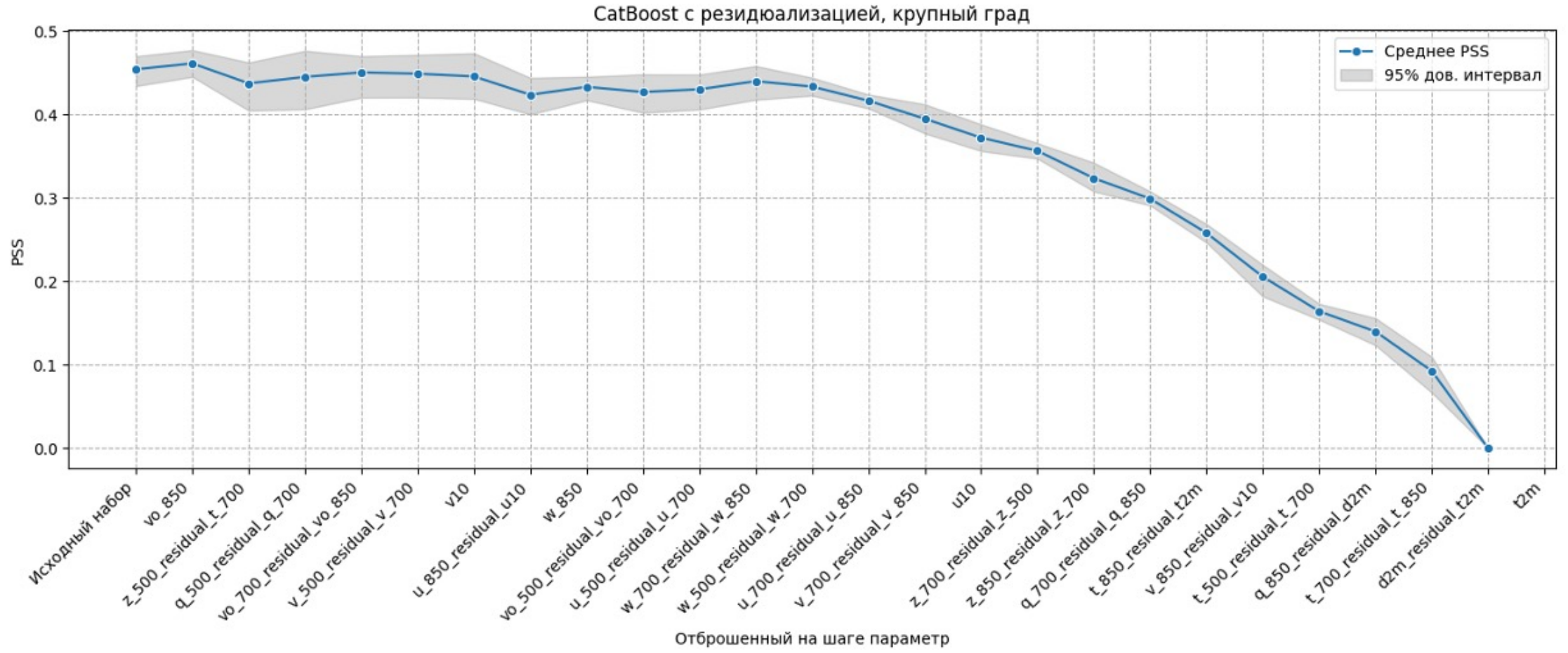


Резидюализация



По цепочке, например: t2m -> t_850 -> t_700 -> t_500

Отбор параметров



Сравнение подходов

	CNN (11x11)	Catboost (только центр)	Catboost (11x11)	Порог WMAXSHEAR>290
PSS	**0.62**	0.50	0.55	0.55
CSI	0.44	0.44	**0.50**	0.27
F1	0.59	0.61	**0.66**	0.41

$$\boldsymbol{PSS} = \frac{TP}{TP + FN} - \frac{FP}{FP + TN}$$

$$\boldsymbol{CSI} = \frac{TP}{TP+FP+FN}$$

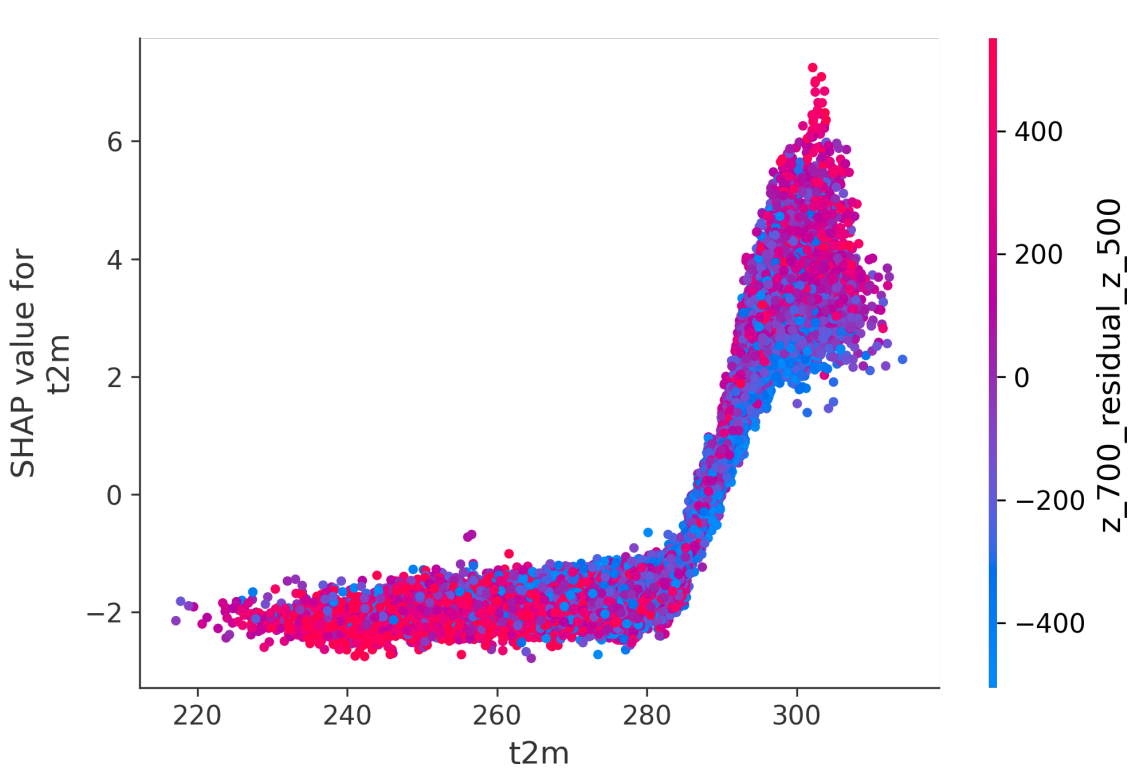
$$\boldsymbol{F1} = \frac{TP}{TP+(FP+FN)/2}$$

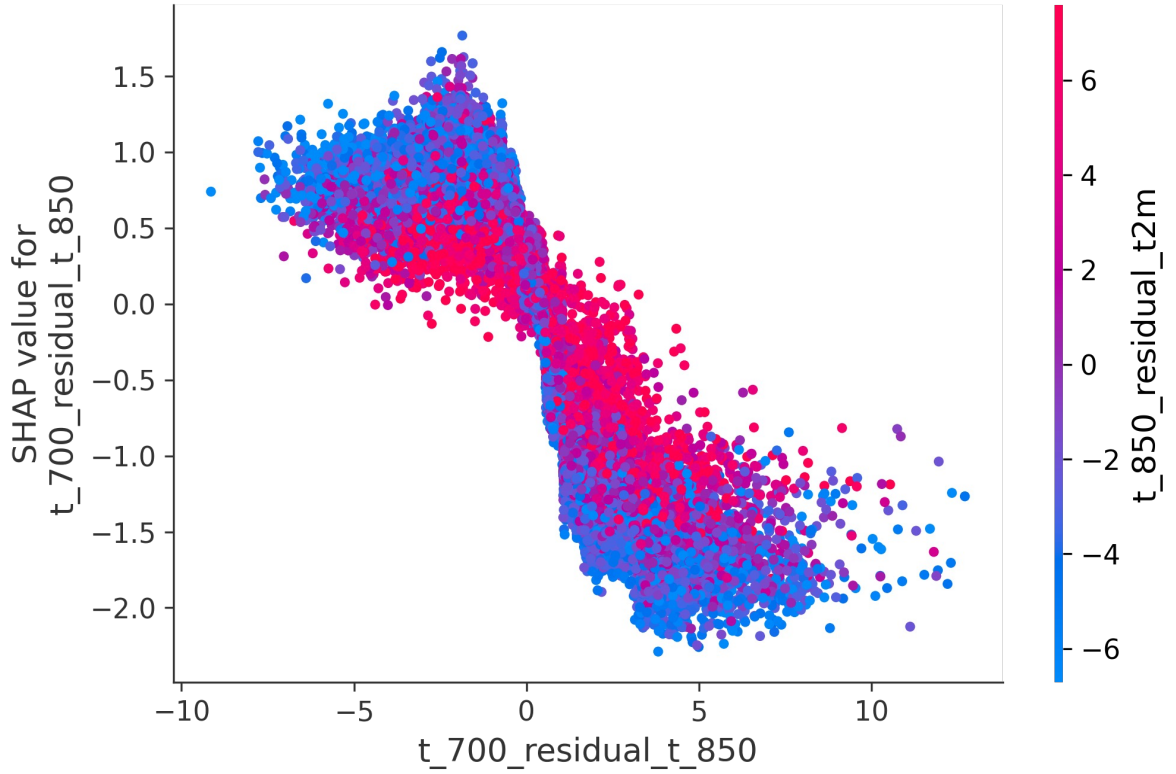
$$\boldsymbol{WMAXSHEAR} = DLS \cdot \sqrt{2 \cdot muCAPE},$$

где $DLS = \sqrt{(u_{500hPa} - u_{10m})^2 + (v_{500\text{hPa}} - v_{10m})^2}$

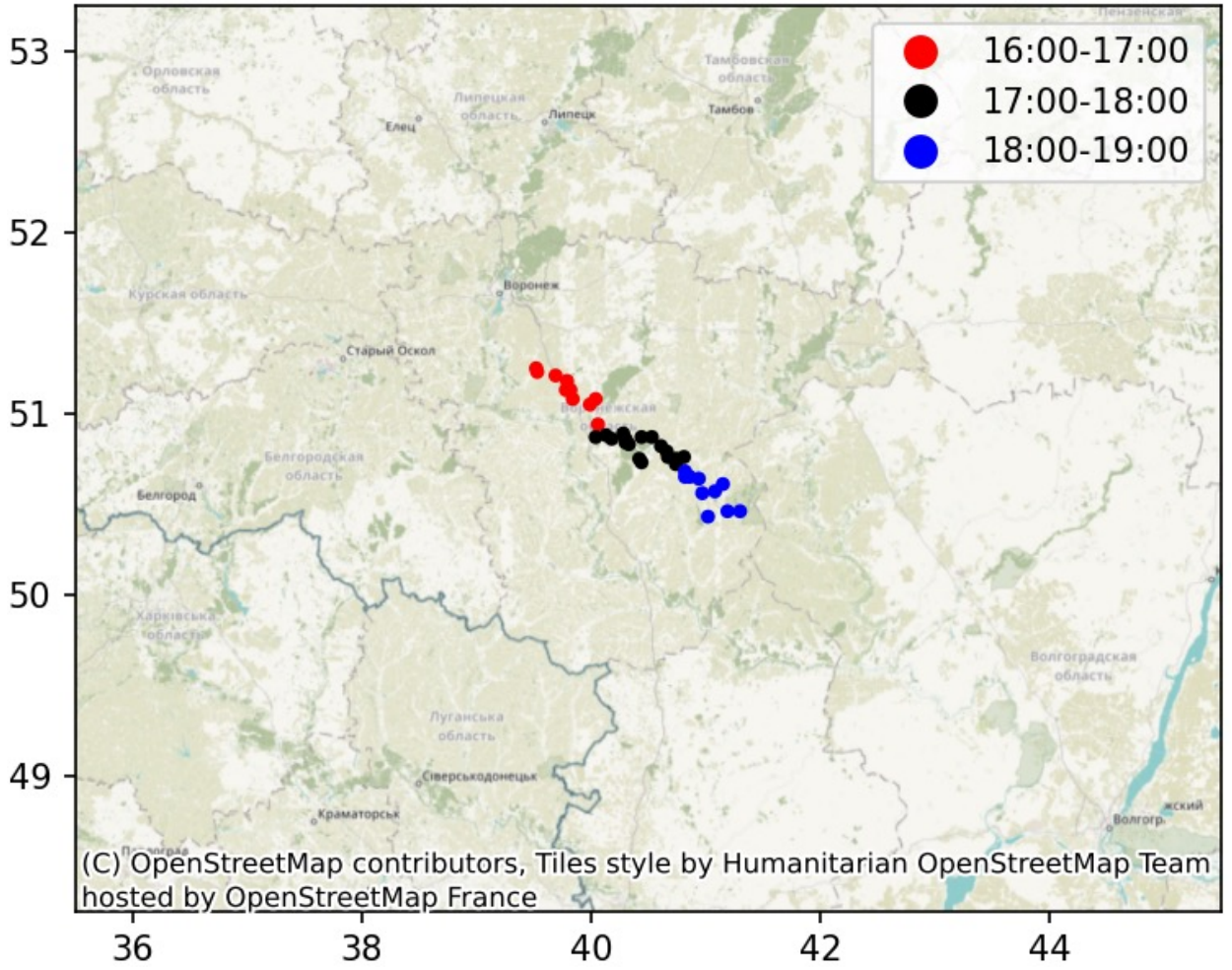
Значение Шэпли – среднее изменение результата при включении параметра в модель по всем возможным подмножествам параметров

$$\boldsymbol{\varphi_i}(v) = \frac{1}{|N|}\sum_{S\subseteq N\setminus\{i\}} \binom{|N|-1}{|S|}^{-1} \left(v(S\cup\{i\}) - v(S)\right)$$

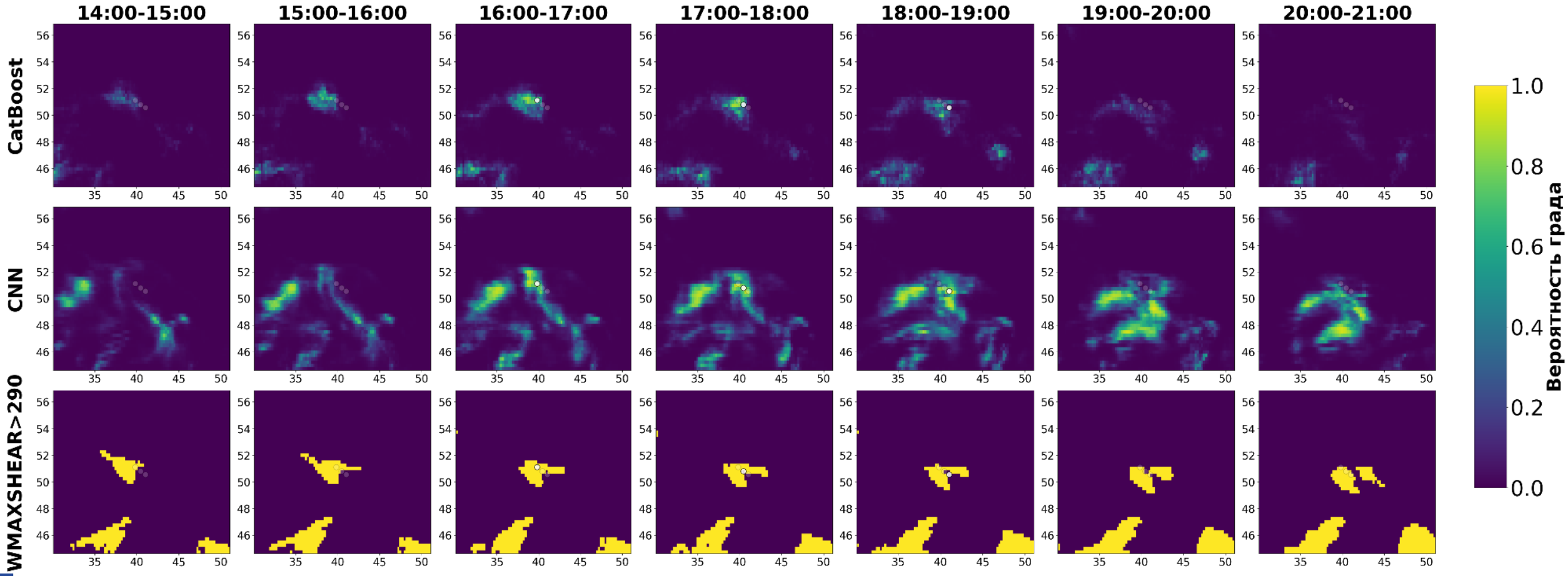




Пример работы моделей: Воронежская область, 24 июня 2024



Пример работы моделей: Воронежская область, 24 июня 2024



Выводы

- Впервые на российских наблюдениях разработана модель МО для диагностики крупного града
- Проведена работа по улучшению интерпретируемости моделей
- Исследованы возможности сокращения объема входных данных
- Проведено изучение поведения моделей на реальном примере

Работа поддержана Российским научным фондом (грант №24-17-00357).

Ограничения

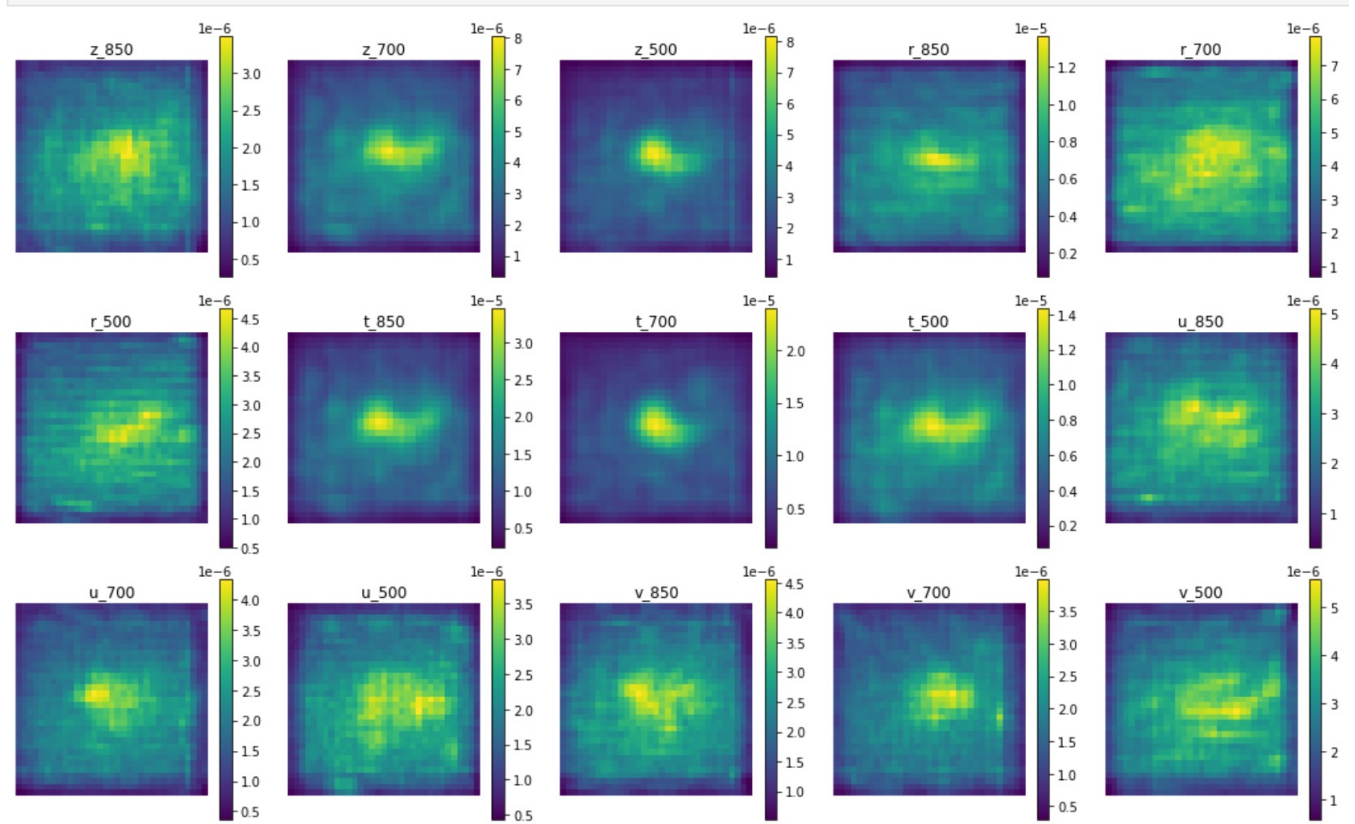
- Нет единого бенчмарка – невозможно сравнить метрики между работами
- Репрезентативность выборки отрицательных событий: будем расширять, добавлять ливни
- Нужны метрики, учитывающие пространственно-временную локализацию

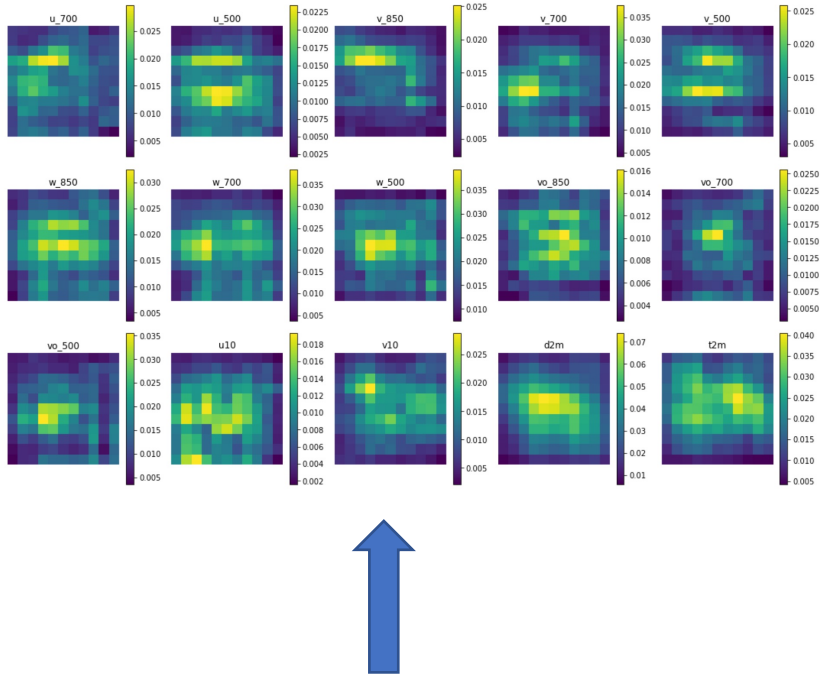
Дальнейшее развитие

- Тестирование применимости в оперативных прогнозах и климатических моделях
- Калибровка вероятностей
- Использование обученной CNN для кластеризации событий по условиям формирования

Предварительный выбор размера окна

Saliency map: **Сильная концентрация значимости в центре**

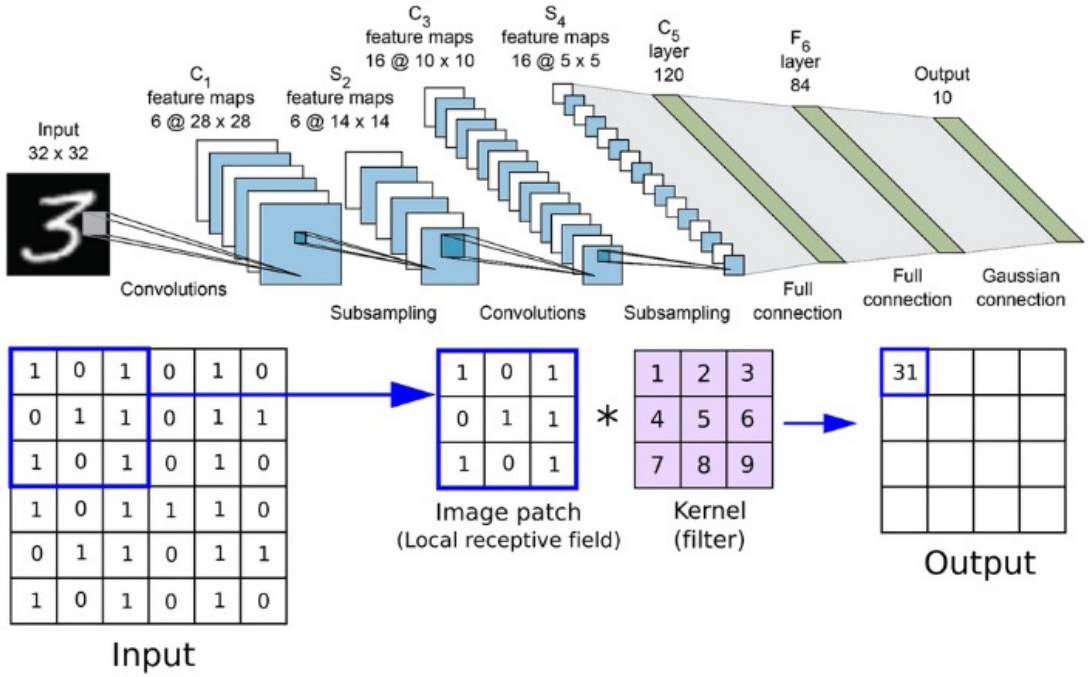




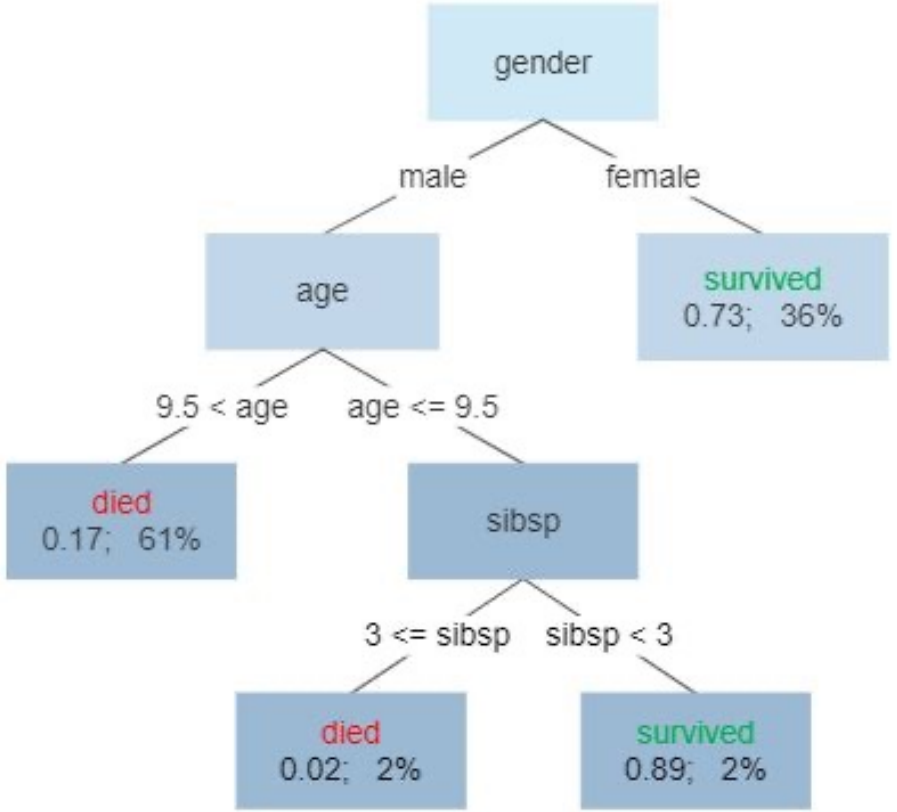
Уменьшили окно до **11x11 пикс. (2.75°x2.75°)**

Модели: общая информация

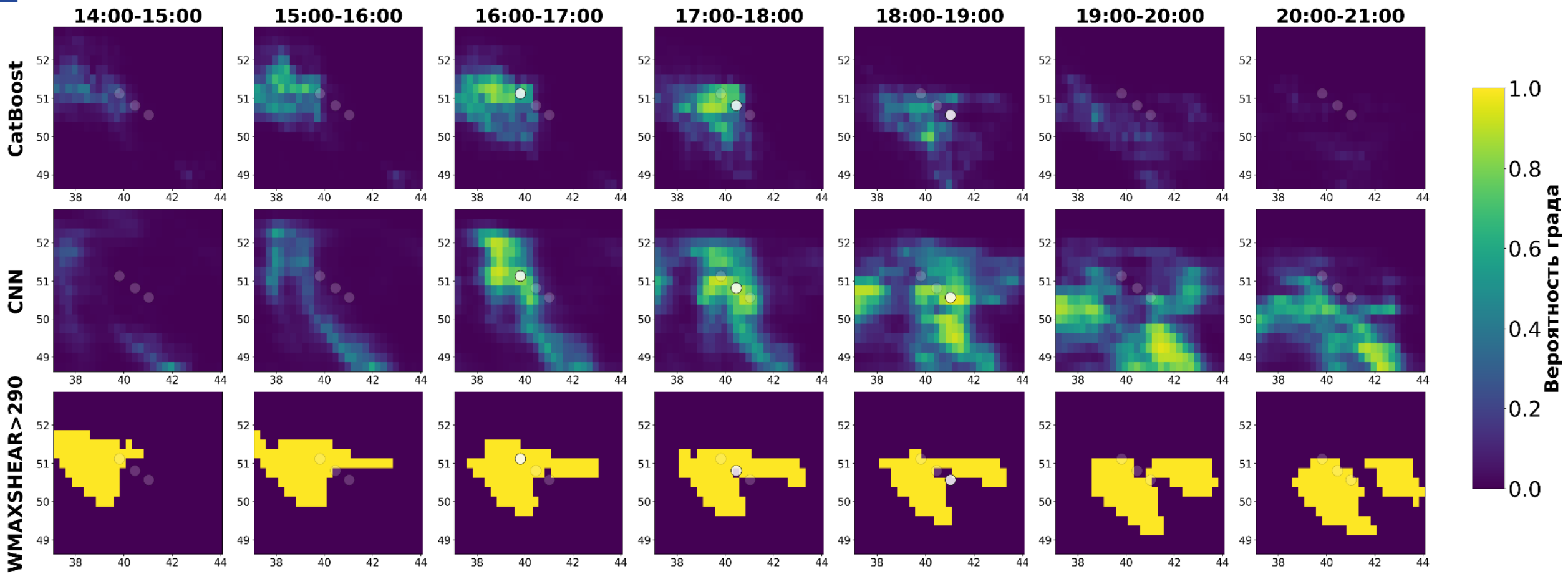
CNN



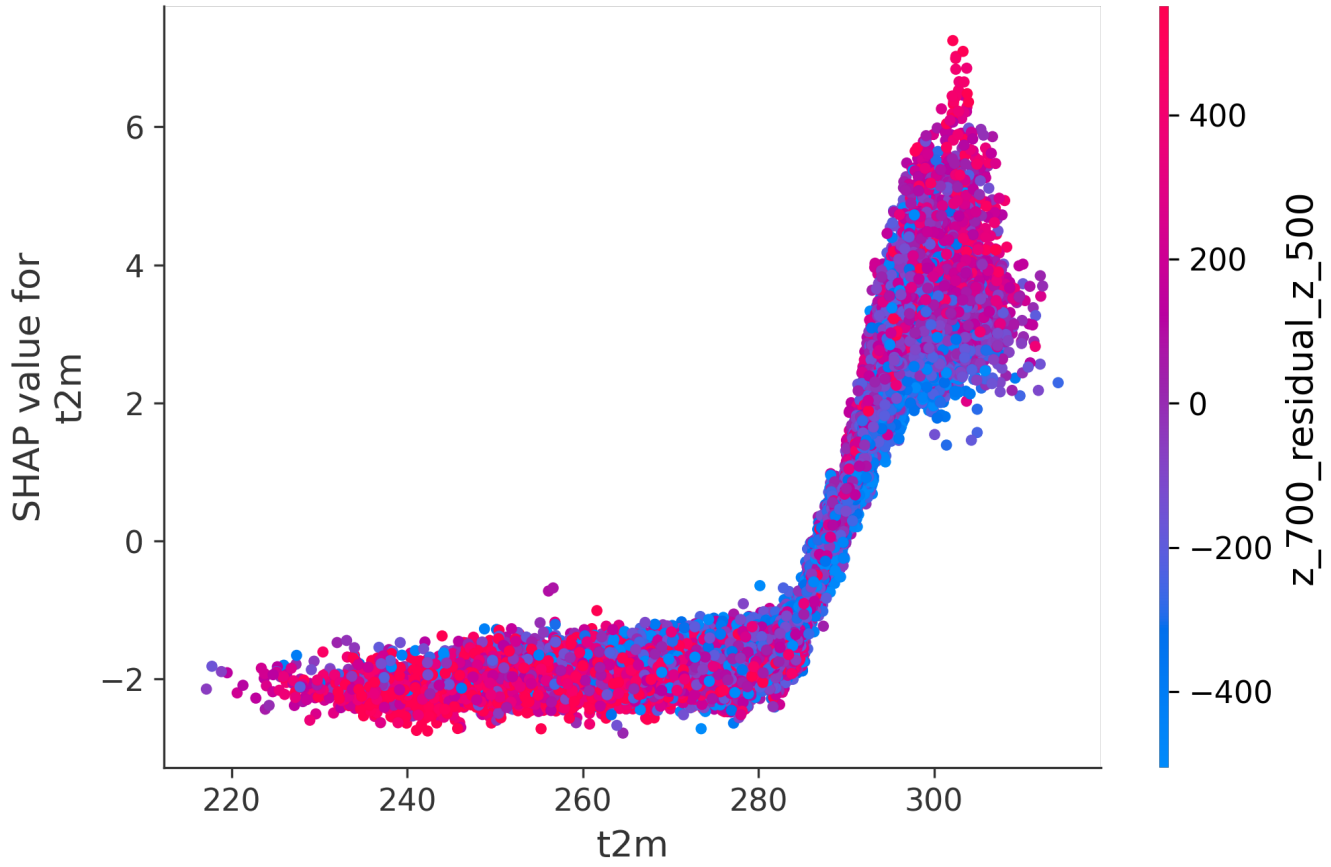
Решающие деревья

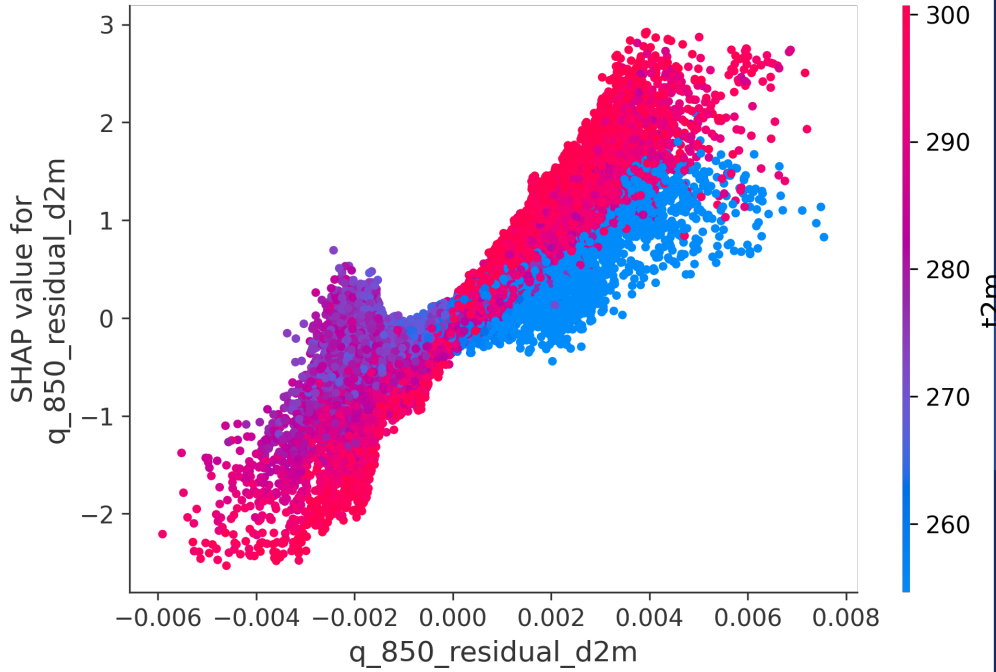


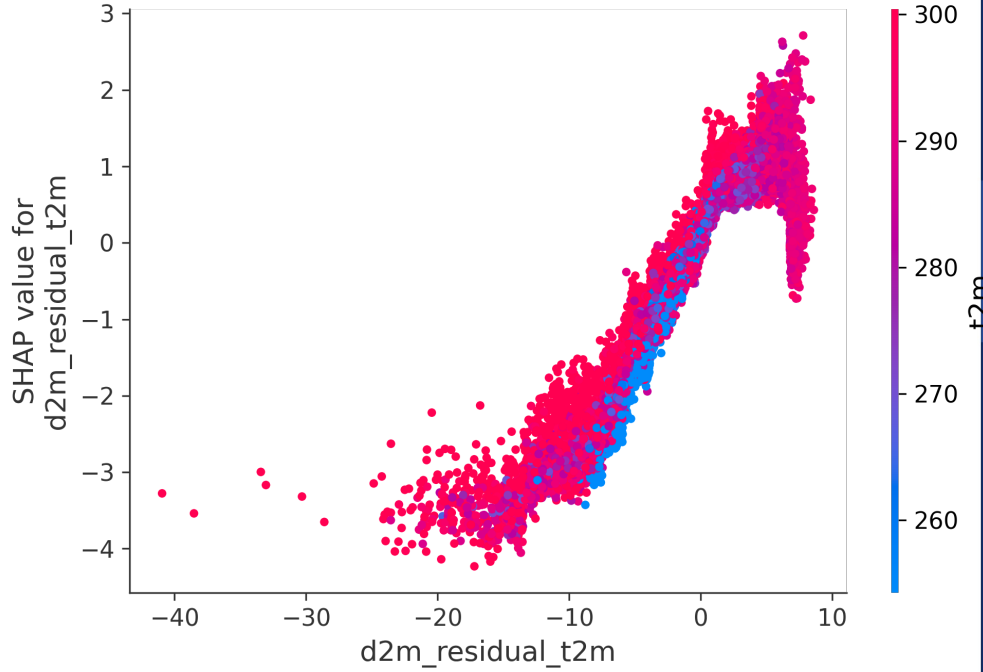
Пример работы моделей: Воронежская область, 24 июня 2024



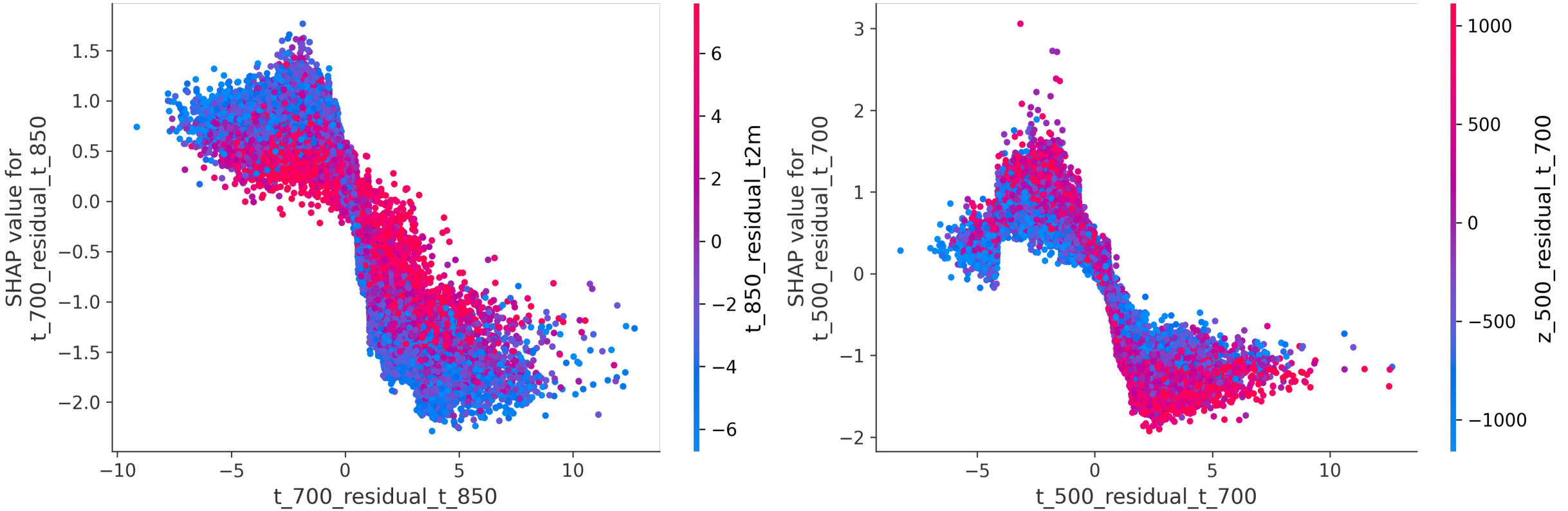
Теплый и влажный воздух внизу



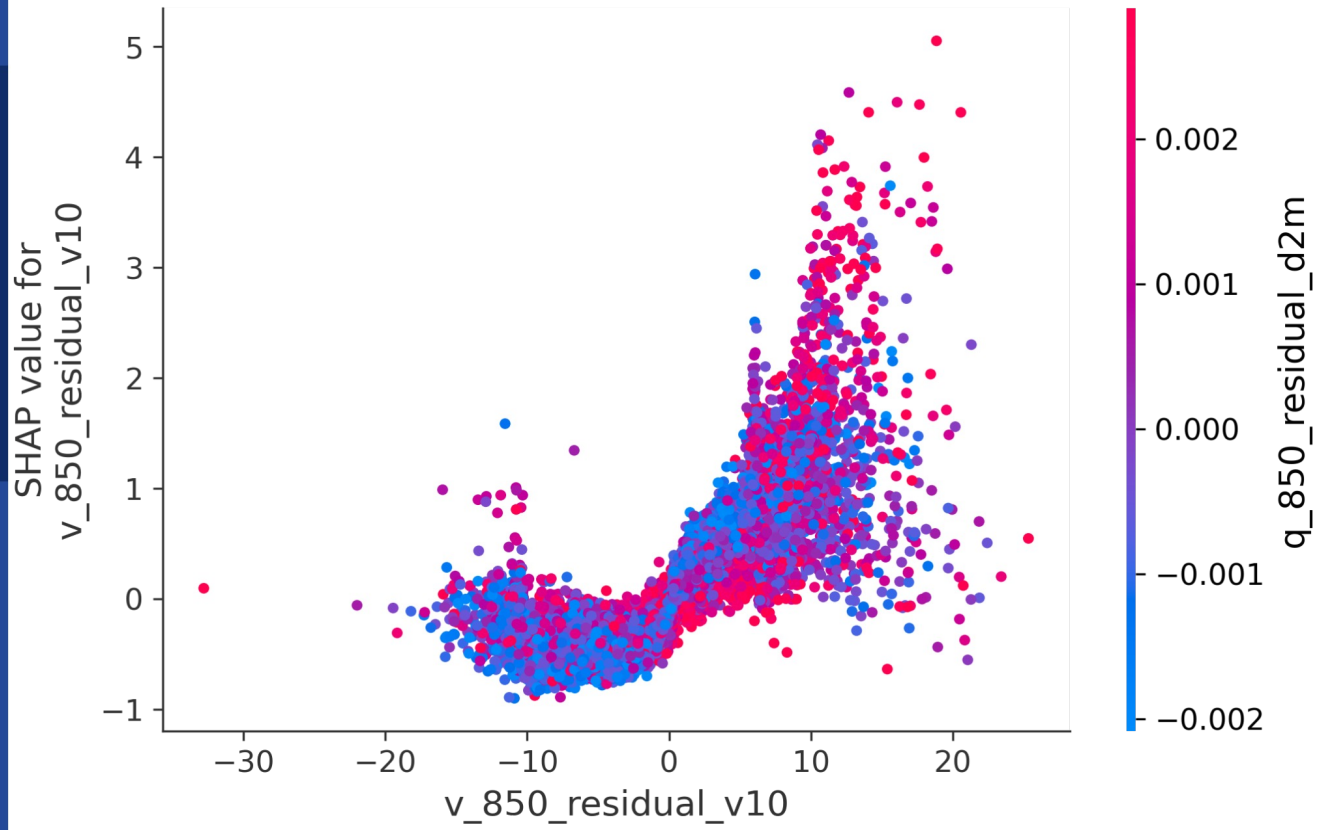


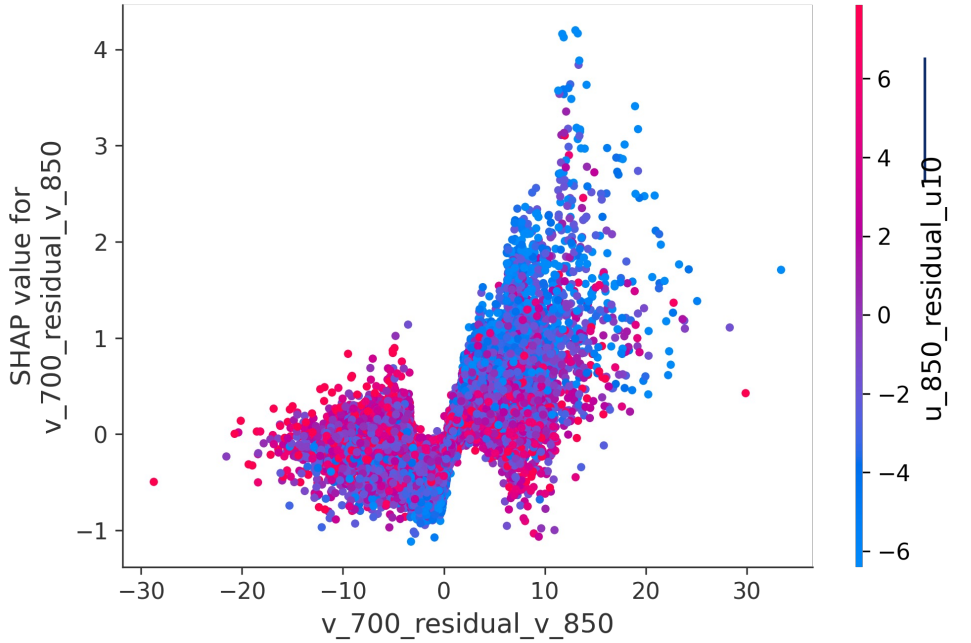


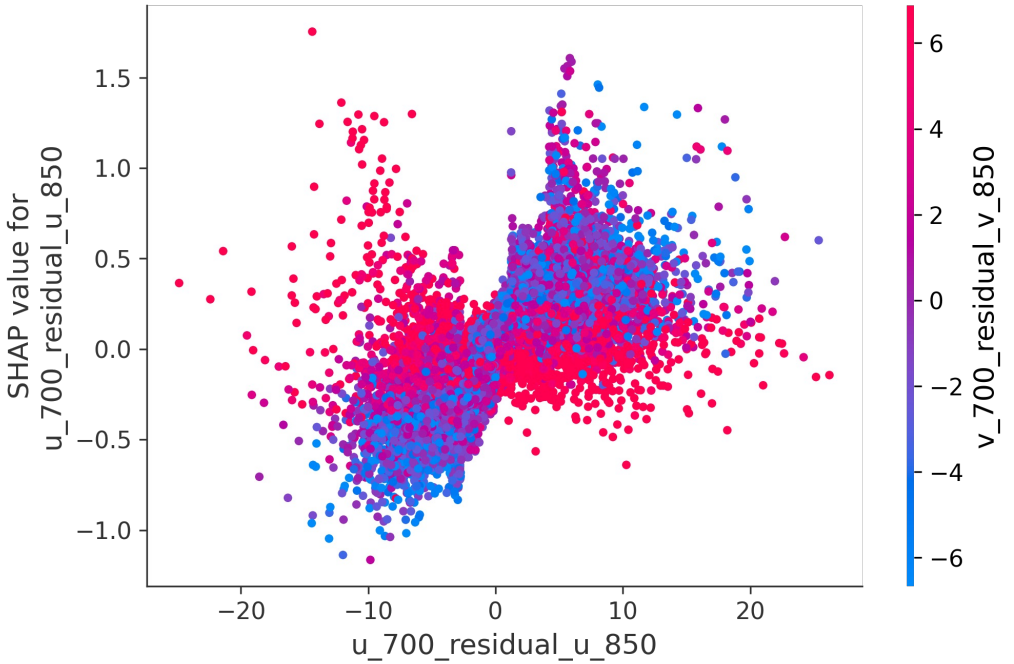
Неустойчивый профиль температуры



Сдвиг ветра


SHAP value for v_850_residual_v10
v_850_residual_v10
q_850_residual_d2m
0.002
0.001
0.000
−0.001
−0.002


SHAP value for v_700_residual_v_850
v_700_residual_v_850
u_850_residual_u10


SHAP value for u_700_residual_u_850
u_700_residual_u_850
v_700_residual_v_850