



# Восстановление приповерхностной влажности атмосферы над океаном с применением методов машинного обучения

*С. А. Вострикова<sup>1</sup>*

*М. А. Криницкий<sup>1,2</sup>*

*М. П. Алекандрова<sup>2</sup>*

*С. К. Гулев<sup>2</sup>*

*<sup>1</sup>Московский физико-технический институт*

*<sup>2</sup>Институт океанологии им. П. П. Ширшова РАН*

# Актуальность

---

- Влажность воздуха — ключевой климатический параметр
- **Измерения влажности разрежены** в пространстве и времени, особенно в начале 20 века
- Охват измерений других метеорологических величин более полный

# Актуальность



- Восстановление данных позволит продлить климатические ряды и повысить точность анализа
- Существующие методы восстановления недостаточно эффективны – необходимы новые подходы

# Цель и задачи исследования

---

**Цель:** ретроспективно восстановить ряды измерений относительной влажности над океаном

## Задачи:

- 1) формирование коллекции данных для обучения и оценки моделей
- 2) оптимизация и обучение алгоритмов машинного обучения
- 3) оценка качества и сравнение моделей

# Данные

**Источник данных:** DISO3 (Dataset of In-Situ Observations over the Ocean), предобработанное подмножество ICOADS (International Comprehensive Ocean-Atmosphere Data Set)

**Период охвата:** 1950-2017 (123 млн наблюдений)

**Тренировочная выборка:** 70% данных

**Тестовая выборка:** 30% данных

**Признаки:**

-температура атмосферы

-температура поверхности океана

-код погоды согласно кодам ВМО

скорость и направление ветра

- типы и баллы облачности по трем ярусам

атмосферное давление

локальная высота солнца

**Целевая переменная:** относительная влажность

# Метод решения

---

## **Протестированы модели машинного обучения:**

- линейная регрессия
- случайный лес
- градиентный бустинг (CatBoost)
- полносвязная искусственная нейронная сеть

## **Повышена пространственно-временная специфичность:**

- данные разделены на ячейки 5x5 градусов по широте и долготе (10x20 градусов в Южном океане)
- каждая ячейка проанализирована отдельно внутри каждого сезона

# Этапы решения задачи

1

## Оптимизация гиперпараметров:

использована библиотека байесовской оптимизации Optuna

2

## Обучение моделей:

- линейная регрессия
- случайный лес
- полносвязная нейросеть
- CatBoost

3

## Оценка качества моделей:

- количественная оценка (RMSE)
- сравнение с тривиальной моделью

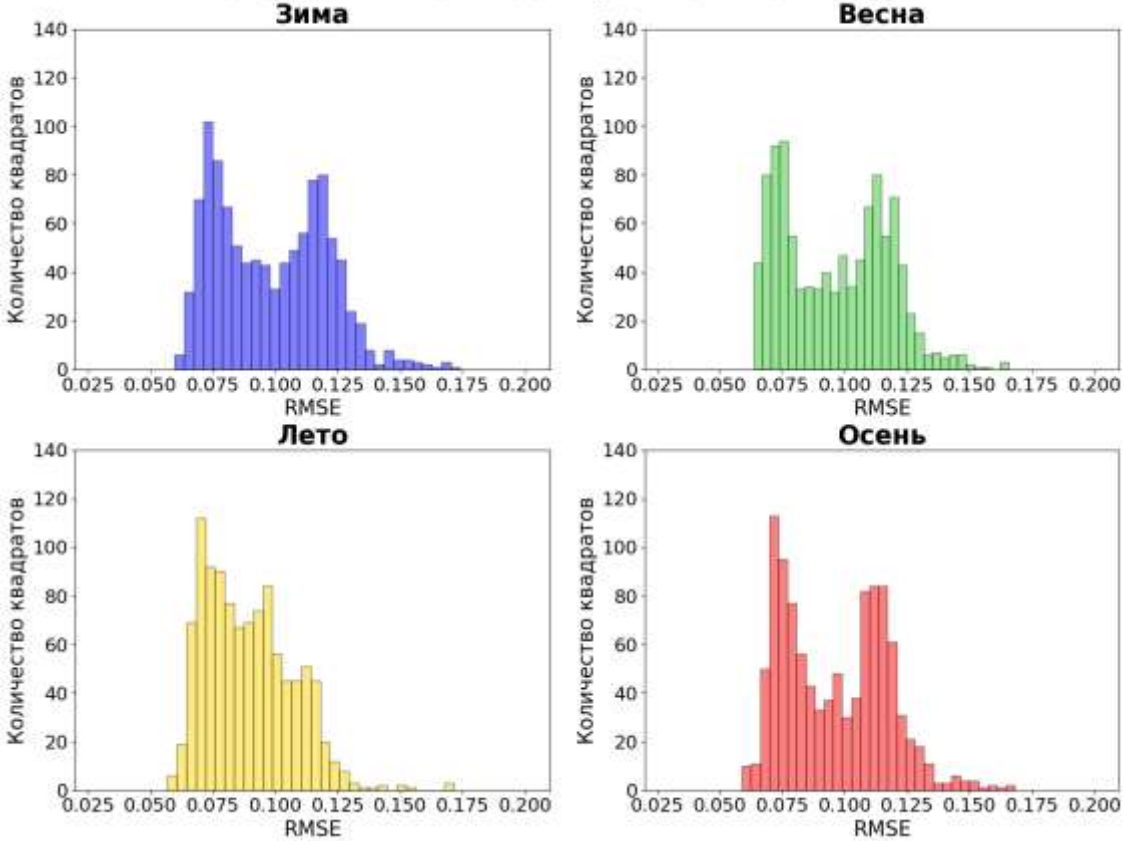
4

## Анализ пространственного распределения точности:

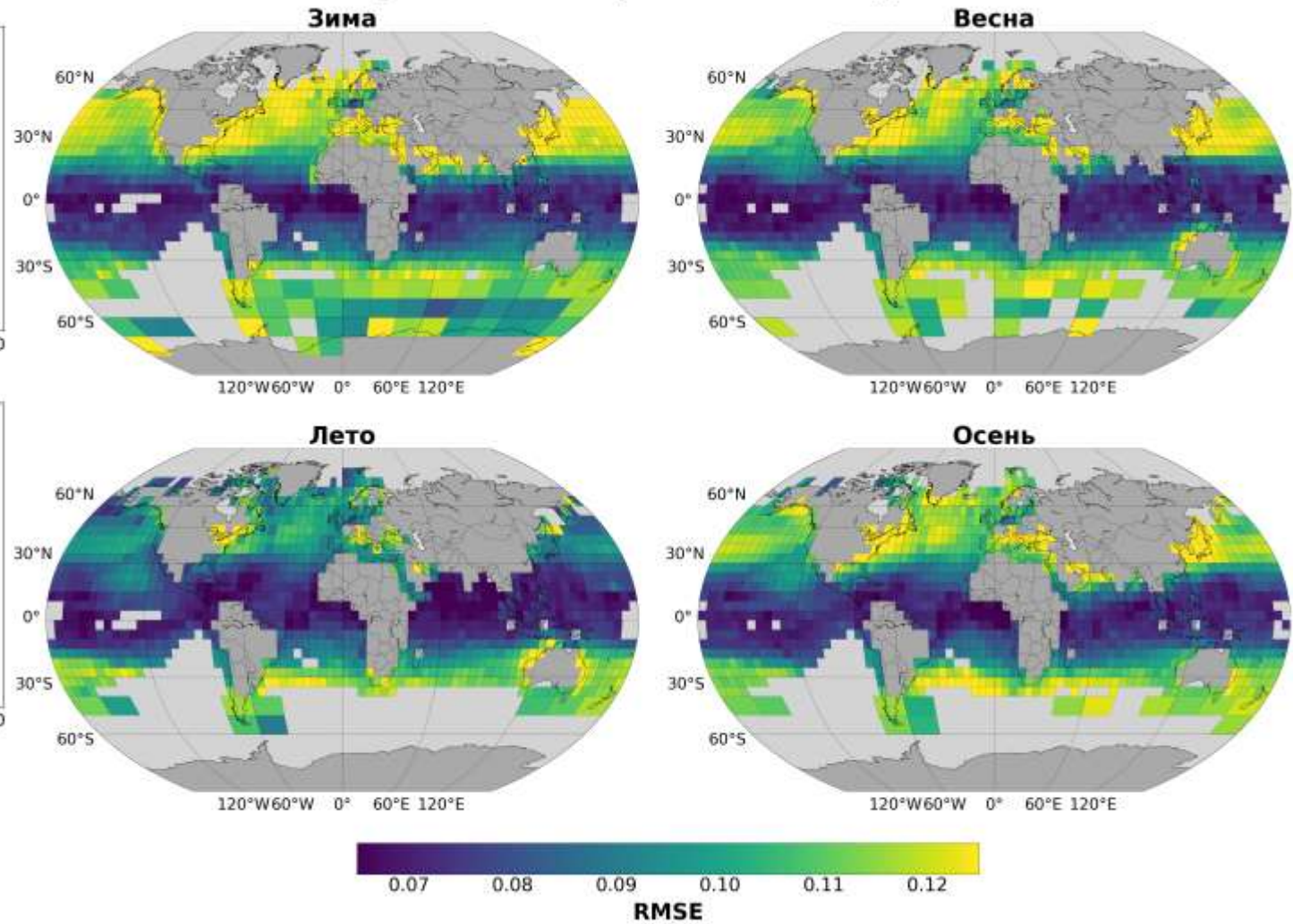
построение и анализ пространственных карт ошибок

# Результаты: тривиальная модель

Гистограммы ошибок тривиальной модели

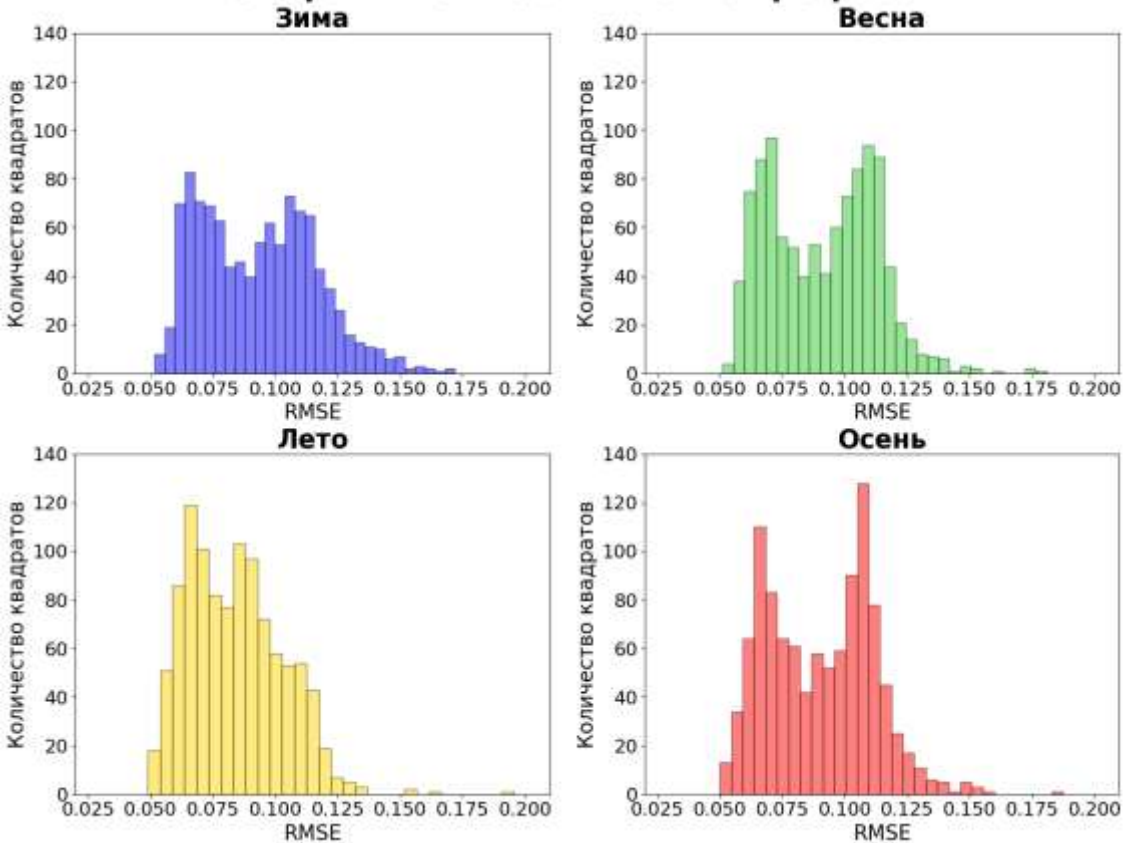


Карты ошибок тривиальной модели

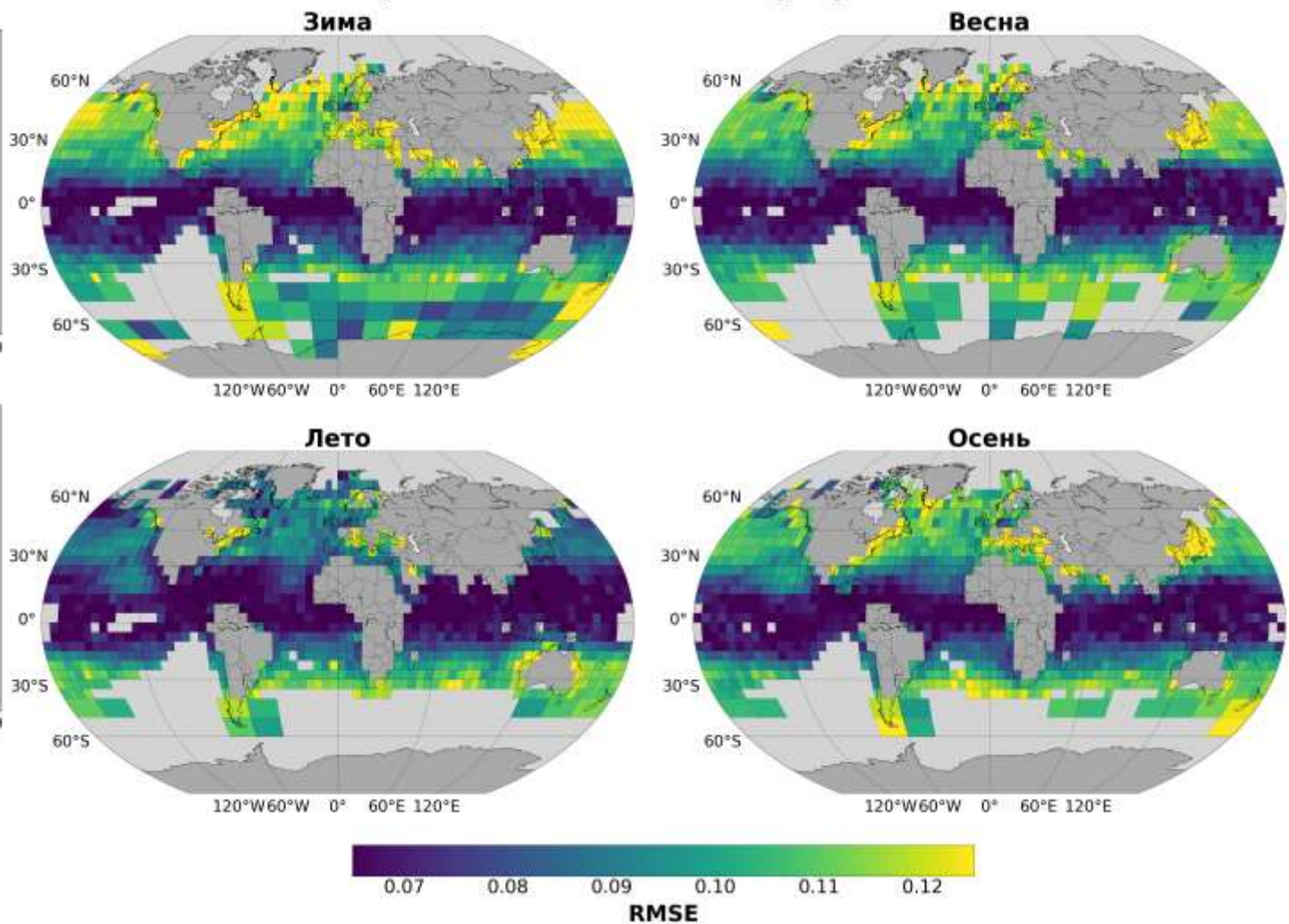


# Результаты: линейная регрессия

## Гистограммы ошибок линейной регрессии

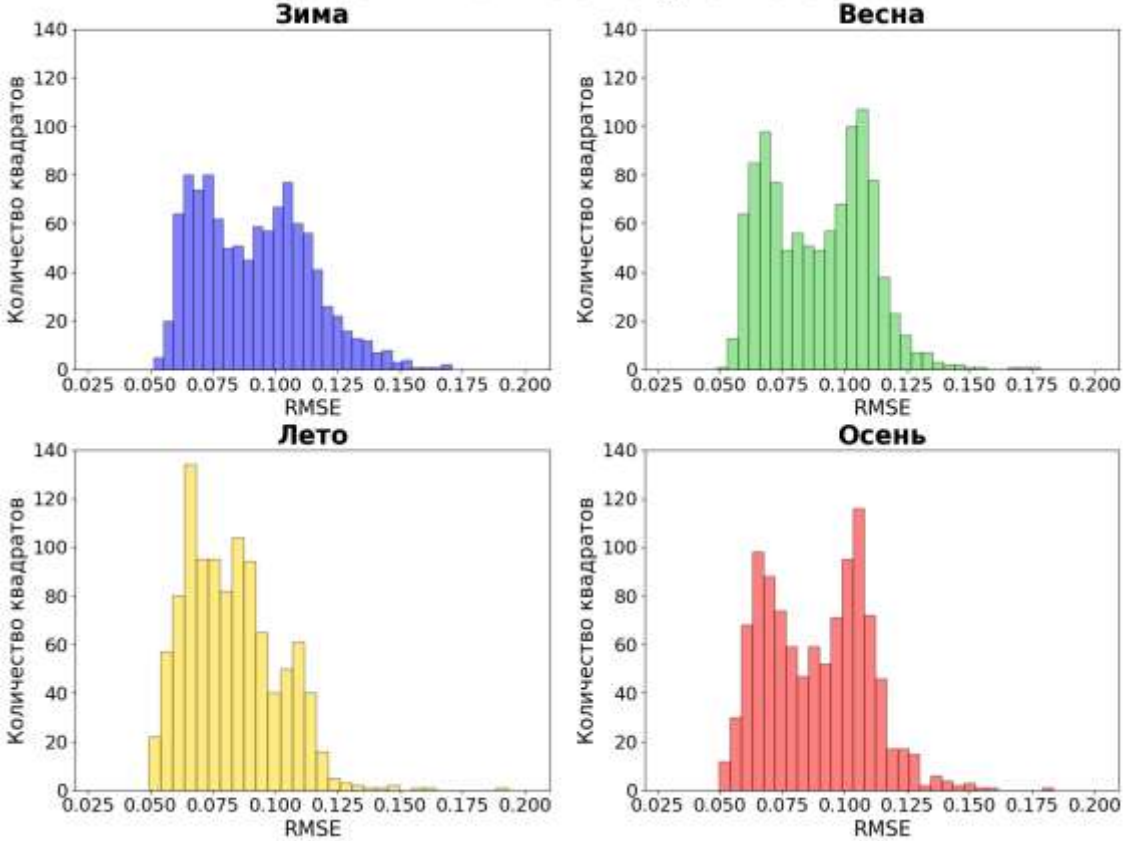


## Карты ошибок линейной регрессии

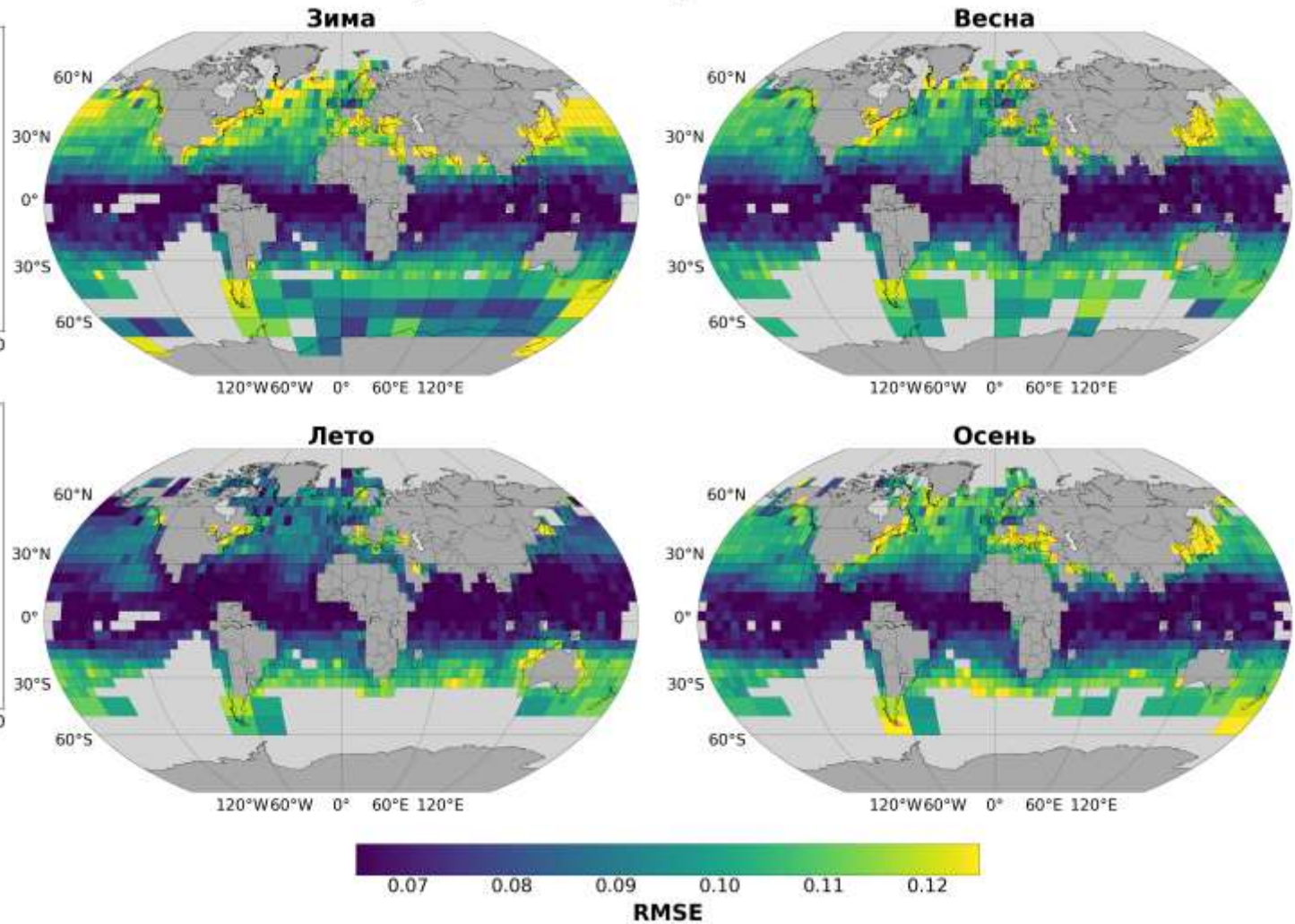


# Результаты: случайный лес

## Гистограммы ошибок случайного леса

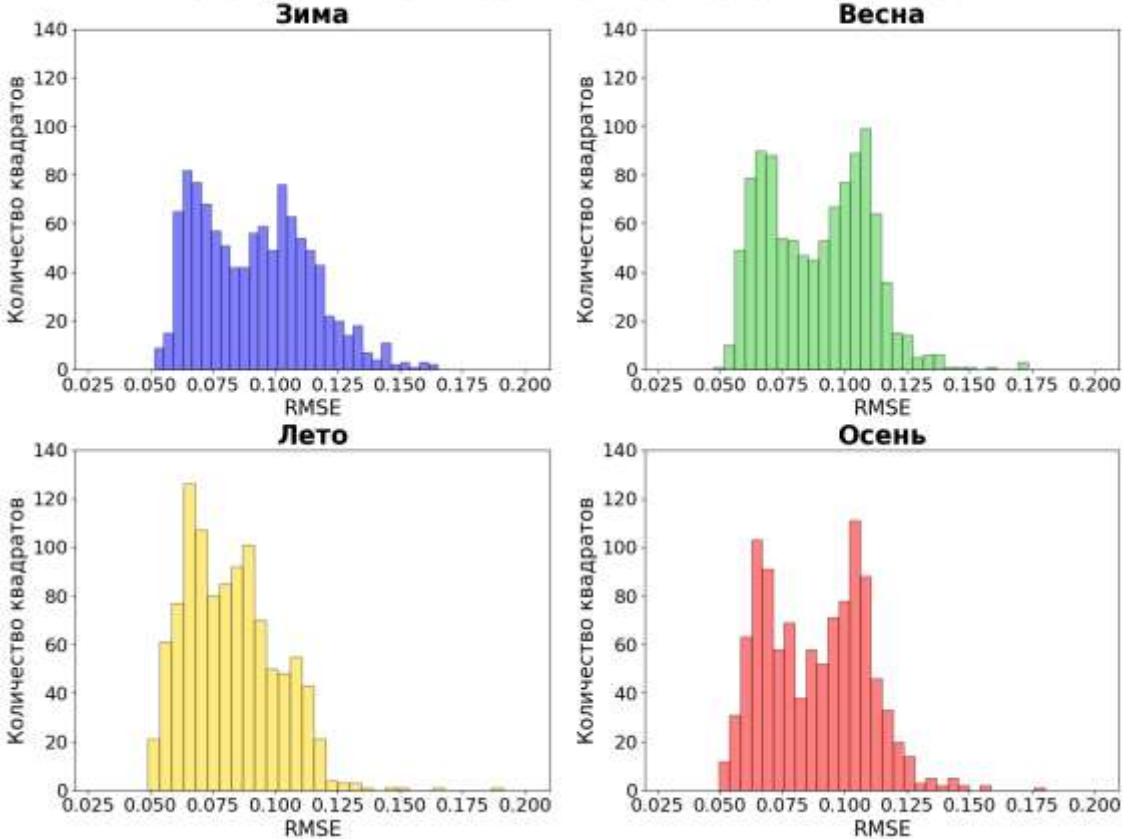


## Карты ошибок случайного леса

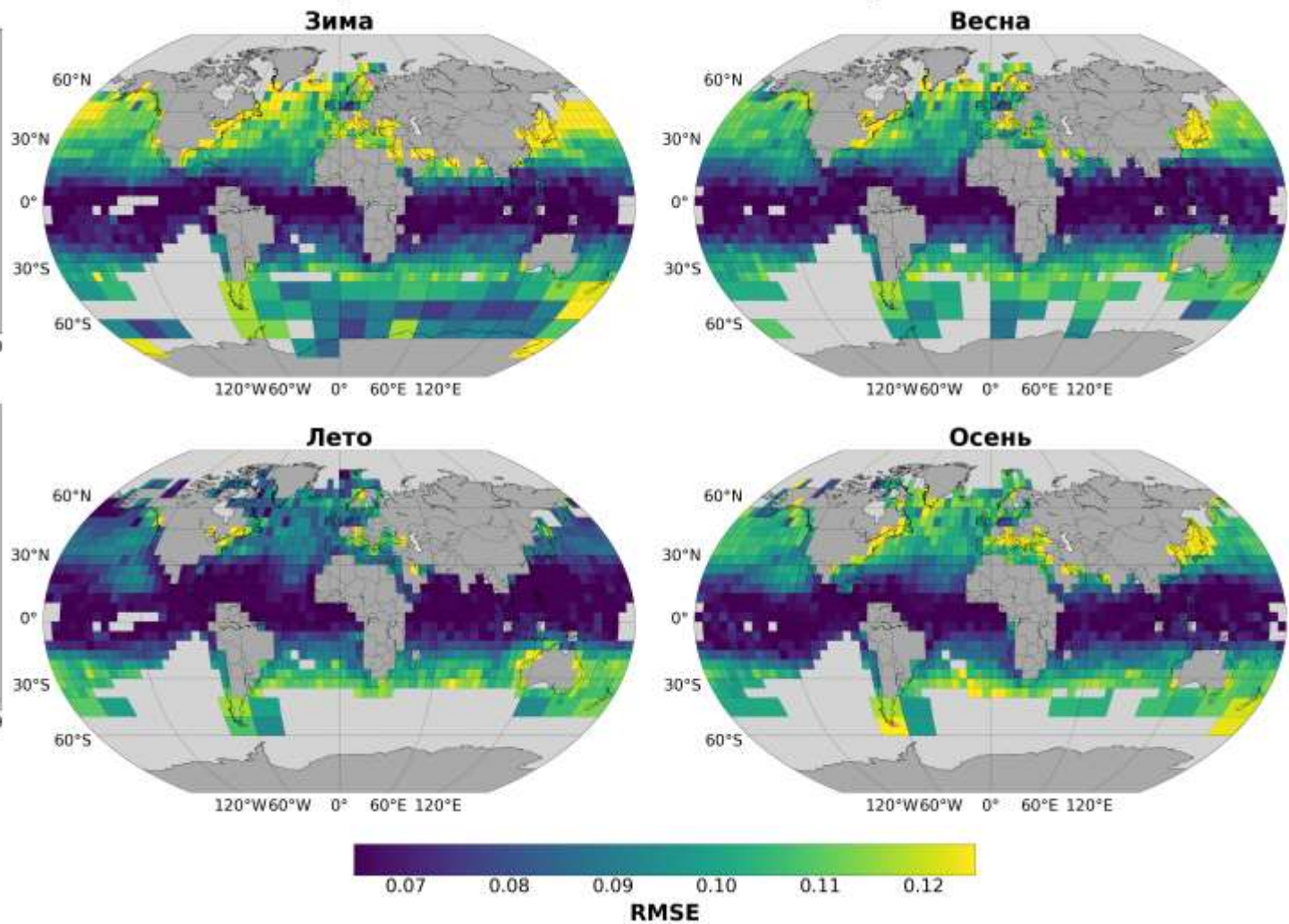


# Результаты: полносвязная нейросеть

Гистограммы ошибок полносвязной нейросети

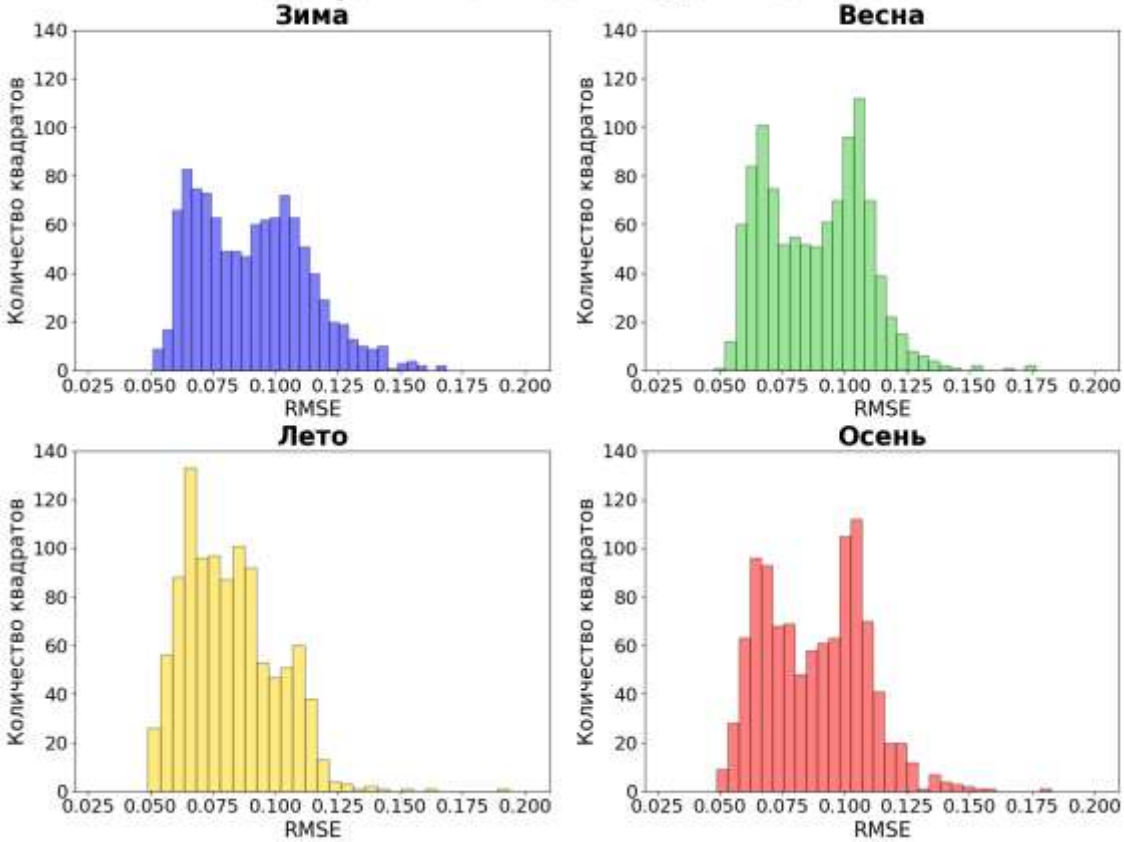


Карты ошибок полносвязной нейросети

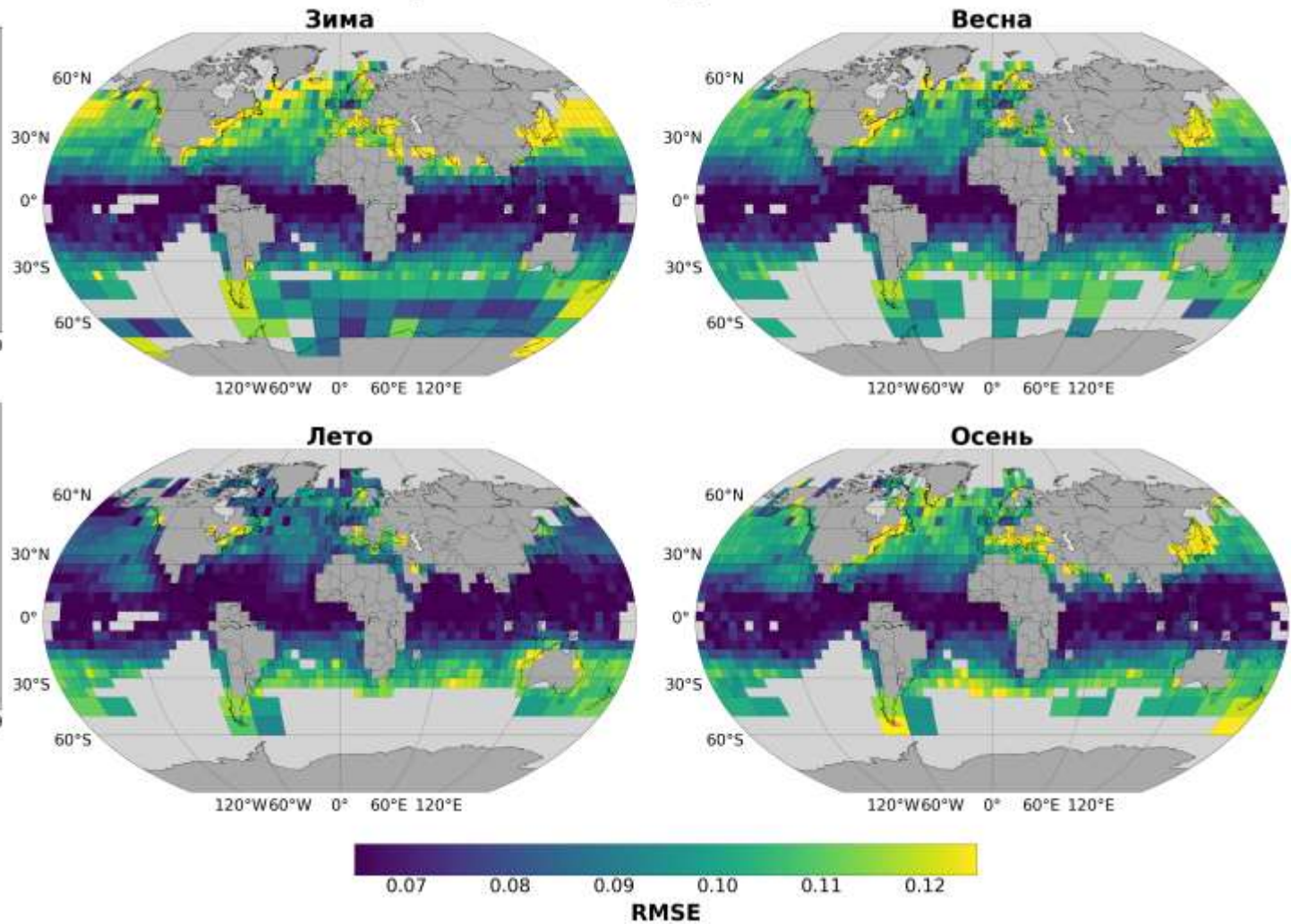


# Результаты: CatBoost

## Гистограммы ошибок модели CatBoost



## Карты ошибок модели CatBoost



# Результаты: качество на отложенной выборке

| Модель                 | Среднее (по океану) RMSE (в долях) |        |        |        |
|------------------------|------------------------------------|--------|--------|--------|
|                        | Зима                               | Весна  | Лето   | Осень  |
| Тривиальная модель     | 0.0981                             | 0.0962 | 0.0892 | 0.0961 |
| Линейная регрессия     | 0.0932                             | 0.0912 | 0.0838 | 0.0904 |
| Случайный лес          | 0.0916                             | 0.0894 | 0.0824 | 0.0887 |
| Полносвязная нейросеть | 0.0914                             | 0.0891 | 0.0821 | 0.0884 |
| CatBoost               | 0.0906                             | 0.0883 | 0.0818 | 0.0881 |

- **CatBoost** – лучшая модель по среднему качеству на отложенной для теста выборке

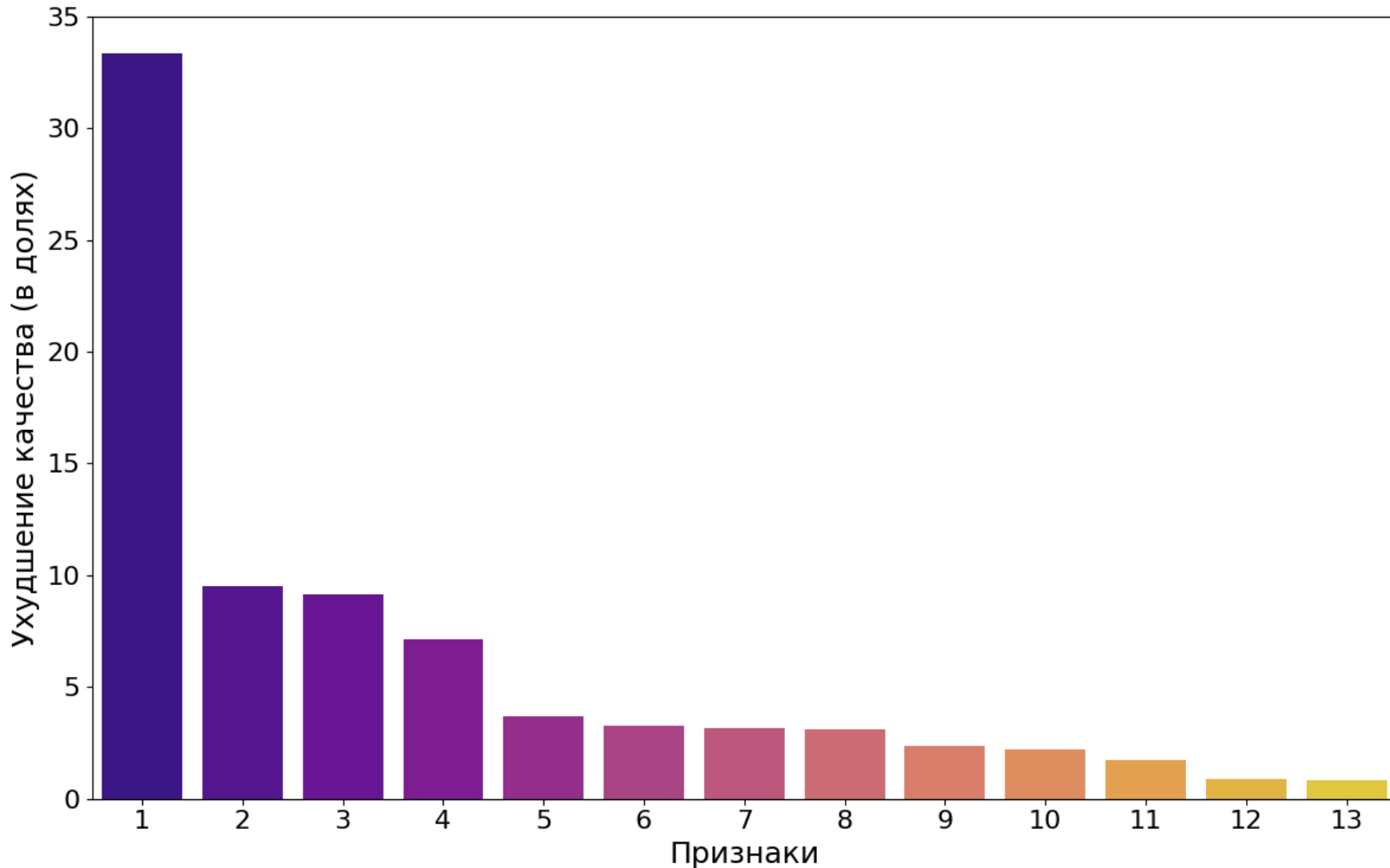
# Результаты: сравнение с тривиальной моделью

| Модель                 | Относительная ошибка (в %) |       |      |       |
|------------------------|----------------------------|-------|------|-------|
|                        | Зима                       | Весна | Лето | Осень |
| Линейная регрессия     | 5.31                       | 5.37  | 6.08 | 5.90  |
| Случайный лес          | 6.78                       | 7.02  | 7.57 | 7.46  |
| Полносвязная нейросеть | 7.07                       | 7.33  | 7.96 | 7.81  |
| CatBoost               | 7.80                       | 8.15  | 8.21 | 8.05  |

- **CatBoost** существенно лучше тривиальной модели

# Значимость признаков для модели CatBoost (permutation importance)

Суммарная (по квадратам) значимость признаков для модели CatBoost



- 1 – температура атмосферы
- 2 – косинус направления ветра
- 3 – локальная высота солнца
- 4 – атмосферное давление
- 5 – код погоды
- 6 – балл нижней облачности
- 7 – температура поверхности океана
- 8 – балл общей облачности
- 9 – синус направления ветра
- 10 – скорость ветра
- 11 – тип нижней облачности
- 12 – тип средней облачности
- 13 – тип верхней облачности

- Модель согласуется с характерными для атмосферы зависимостями

# Оценка неопределённостей для модели CatBoost (bootstrap)

| Сезон | Средняя (по квадратам) неопределённость для модели CatBoost (в долях) |                                     |
|-------|-----------------------------------------------------------------------|-------------------------------------|
|       | Неопределённость меры качества                                        | Неопределённость целевой переменной |
| Зима  | 0.00049                                                               | 0.0106                              |
| Весна | 0.00048                                                               | 0.0102                              |
| Лето  | 0.00059                                                               | 0.0109                              |
| Осень | 0.00061                                                               | 0.0122                              |

- Модель работает устойчиво и даёт надёжные результаты, даже при изменении входных данных

# Выводы

---

- Алгоритмы машинного обучения показали высокую эффективность при восстановлении влажности
- Достигнуто значительное улучшение точности по сравнению с тривиальным подходом (усреднением)
- **CatBoost** – лучшая модель для восстановления влажности
- Разработанный подход может использоваться для заполнения пропусков в климатических рядах