МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ПРОГНОЗИРОВАНИЕ ТЕМПЕРАТУРЫ ВОЗДУХА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование
информационных систем
факультета компьютерных наук и информационных технологий
Агаркова Максима Андреевича

Научный руководитель:
зав. кафедрой, к.ф.-м.н., доцент _____ Огнева М.В.
подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М.В.
подпись, дата

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Точность температурного прогноза имеет критическое значение для множества отраслей и направлений хозяйственной деятельности. Оценка риска заморозков для сельского хозяйства, расчёты пиковых нагрузок в энергосистеме, планирование потребности в теплоресурсах — во всех этих случаях ошибка в несколько градусов приводит к существенно иным хозяйственным решениям. При этом большинство доступных метеорологических продуктов и моделей погоды дают оценку не в конкретной точке наблюдения, а на регулярной сетке, усредняя состояние атмосферы над площадью в сотни квадратных километров. Для приведения таких оценок к уровню конкретных пунктов наблюдений применяются методы статистической постобработки прогноза.

Реанализ ERA5 [1] — это оценка состояния атмосферы на регулярной сетке с шагом около 31 км. Когда нужно сопоставить его данные с конкретной метеостанцией, возникает практически важная проблема: ячейка сетки может включать городской квартал, открытую степь и водоём, тогда как станция измеряет температуру в одной точке. Разница между значением в узле сетки и показанием станции бывает устойчивой — она не исчезает при усреднении по многим дням, а значит, систематически искажает оценку температурного режима.

Параллельно существует спутниковый продукт MODIS LST — температура поверхности с разрешением один километр. Это другая физическая величина: температура поверхности и температура воздуха на высоте 2 метров ведут себя по-разному при облачности, влажности, сезоне и времени суток. Вдобавок MODIS регистрирует показания только при ясном небе — в саратовском наборе пропуски по дневной и ночной температуре поверхности составили около 61 и 60 процентов соответственно [2].

Совместное использование реанализа ERA5, спутниковых наблюдений MODIS и наземных станционных данных позволяет построить модель машинного обучения, которая восстанавливает температуру воздуха точнее,

чем любой из этих источников по отдельности — это подтверждается в исследованиях для Китая [3] и Ганы [4]. При работе с табличными метеорологическими данными особую эффективность показывают ансамблевые методы на деревьях решений, в частности градиентный бустинг XGBoost.

Цель бакалаврской работы – разработать и проверить воспроизводимую схему машинного обучения для восстановления среднесуточной температуры воздуха по данным ERA5, MODIS и наземных станций Росгидромета, а также оценить применимость схемы к задаче калибровки открытых источников температуры относительно официальных наблюдений.

Поставленная цель определила следующие задачи:

1. Выполнить обзор существующих методов прогнозирования температуры воздуха и способов оценки данных методов.

2. Подготовить объединённый набор данных ERA5, MODIS и наземных станций, описать его ограничения и структурные пропуски.

3. Построить и сравнить простые базовые модели, чтобы обосновать выбор XGBoost как основной модели.

4. Реализовать модель на основе градиентного бустинга и проверить, как на качество влияют лаговые, пространственные и календарные признаки.

5. Оценить устойчивость модели по месяцам, станциям и временным периодам.

6. Проверить переносимость построенной схемы на Волгоградскую область.

7. Разработать и оценить калибровку RP5-подобного источника к данным Росгидромета, включая оценку интервалов неопределённости.

В работе использованы исследования в области реанализа ERA5 [1], алгоритма XGBoost [5] и ансамблевых методов на деревьях решений [6], а также метода конформного оценивания неопределённости [7].

Теоретическая и практическая значимость бакалаврской работы. Научная значимость — систематическая проверка методов коррекции сеточных данных реанализа на материале российских метеостанций с контролем временно́й последовательности обучения и тестирования. Практическая значимость — методика уточнения открытых источников температурных данных по официальным наблюдениям Росгидромета, пригодная для использования в прикладных задачах.

Структура и объём работы. Бакалаврская работа состоит из введения, 2 разделов, заключения, списка использованных источников и 5 приложений. Общий объем работы – 85 страниц, из них 53 страниц – основное содержание, включая 24 рисунков и 12 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 25 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Постановка задачи и методы» посвящён теоретическим основам работы.

Задача поставлена как регрессионное восстановление среднесуточной температуры воздуха по признакам реанализа ERA5, спутниковых данных MODIS и характеристик пунктов наблюдений. В качестве входных данных используются значения за текущий и три предыдущих дня; значения самой целевой переменной в признаки не включаются, что исключает подмену прогноза ретроспективным усреднением.

Выполнен обзор методов статистической постобработки прогнозов погоды: от простой линейной корректировки смещения до нейросетевых ансамблей. Для задач на табличных данных ансамблевые методы на деревьях решений стабильно превосходят глубокое обучение [6]. Рассмотрены алгоритмы случайного леса, LightGBM, CatBoost и XGBoost.

Алгоритм XGBoost [5] последовательно строит ансамбль деревьев решений: каждое следующее дерево исправляет ошибки предыдущих.

Пропущенные значения обрабатываются автоматически — при обучении определяется, в какую ветку дерева направлять объекты с пропусками, что особенно важно для спутниковых данных MODIS с пропусками до 60%.

Описаны четыре группы входных признаков: поля ERA5 (температура, ветер, влажность, давление за текущий и три предыдущих дня), данные MODIS (дневная и ночная температура поверхности за те же периоды), координаты станции и календарные переменные (день года в виде тригонометрических компонент). Разделение на обучение (2013–2021) и тест (2022–2023) строго выдерживается по времени.

Качество прогноза оценивалось по трём метрикам: среднеквадратическому отклонению (RMSE), средней абсолютной ошибке (MAE) и коэффициенту детерминации R^2. RMSE сильнее штрафует крупные выбросы, тогда как MAE даёт более интуитивную оценку типичной ошибки в градусах Цельсия. Для проверки устойчивости модели к незнакомым станциям применялась схема последовательного исключения: модель обучалась на всех станциях, кроме одной, и проверялась на исключённой. Дополнительно проводился анализ по подвыборкам с полными, частичными и отсутствующими данными MODIS.

Рассмотрены три способа применения саратовской модели в новом регионе. При первом варианте — без переобучения — готовая модель применяется напрямую к волгоградским данным. При втором — с дообучением — параметры модели корректируются на небольшой выборке из нового региона. При третьем — обучение заново — модель строится полностью на волгоградских данных с тем же набором признаков. Для каждого варианта сохраняется строгое временно́е разбиение.

Описан метод конформной квантильной регрессии [7] для построения доверительных интервалов прогноза. Квантильная регрессия формирует предварительные нижнюю и верхнюю границы, которые затем уточняются по контрольной выборке — так, чтобы доля попаданий совпала с заданным

уровнем (например, 85%) без предположений о форме распределения ошибок.

Рассмотрены показатели важности признаков: вклад в снижение ошибки, охват обучающих примеров в узлах и частота использования. Такой анализ позволяет выявить, не доминирует ли в модели один простой признак — например, среднестанционная температура, что могло бы указывать на подмену прогноза усреднением.

Подчёркнута важность пространственной независимости при тестировании: при стандартной кросс-валидации со случайным перемешиванием соседние по времени наблюдения с одной и той же станции одновременно попадают в обучение и тест, что завышает оценку качества. Схема исключения по станциям воспроизводит реальное применение: модель никогда не видела тестируемую станцию при обучении. Результаты в этом режиме оказываются заметно скромнее, чем при случайной кросс-валидации, но зато честнее отражают обобщающую способность модели.

Второй раздел «Практическая реализация и результаты» посвящён практическому применению разработанной схемы.

Собраны и описаны три набора данных. Основной набор — 52 наземные станции Росгидромета в Саратовской области за 2013–2023 гг. с соответствующими полями ERA5 и пикселями MODIS. Пропуски по дневной и ночной температуре поверхности MODIS составили около 61% и 60% соответственно. Набор для переноса — 12 станций Волгоградской области. Набор для калибровки — 125–132 станции из открытого RP5-подобного источника, совпадающих с архивом Росгидромета.

Сравнение шести базовых алгоритмов (линейная регрессия, полиномиальная, метод k ближайших соседей, дерево решений, случайный лес, градиентный бустинг sklearn) показало: XGBoost без специальных признаков уже даёт RMSE = 1.33 °C, тогда как случайный лес — 2.07 °C, линейная регрессия — 3.45 °C. Это обосновало выбор XGBoost как основной модели.

Поэтапное добавление признаков снизило RMSE: базовая XGBoost (ERA5-признаки) — 1.33 °C, добавление лаговых признаков ERA5 и MODIS (t−1, t−2, t−3) — 1.18 °C, добавление координат станций и устранение утечки через station_mean — 1.15 °C. Основной прирост дали именно лаговые признаки, а не усложнение алгоритма.

На тестовом периоде 2022–2023 гг., где доступна исходная температура ERA5, итоговая модель снизила RMSE с 0.9806 до 0.7882 °C (≈19.6%) и MAE с 0.7437 до 0.5978 °C (≈19.6%) по сравнению с прямым использованием ERA5. Признак station_train_mean_T в итоговую модель не вошёл, что подтверждает: модель не сводится к подстановке среднего по станции.

Анализ устойчивости по месяцам показал: наиболее тяжёлый сезон — зимние месяцы (RMSE в январе 1.85 °C), наилучшее качество — летом. Наиболее проблемная станция — 35108 с нетипичным локальным режимом (RMSE 3.2 °C). Попытка построить отдельную зимнюю модель не дала улучшения качества.

Проверка переносимости на Волгоградскую область показала: готовая саратовская модель без адаптации — RMSE 1.21 °C; дообучение на волгоградских данных — RMSE 0.88 °C; обучение с нуля на целевом регионе — RMSE 0.83 °C и MAE 0.63 °C. Следовательно, переносится не сама модель, а выбранная схема признаков и порядок обучения.

В задаче согласования RP5-подобного источника с наблюдениями Росгидромета реализована коррекция систематической погрешности: по каждой станции вычисляется разность средних значений двух источников за один и тот же период, и эта поправка прибавляется к данным RP5-подобного источника. Поправка применяется только там, где она стабильно снижает ошибку на обучающих данных; в остальных случаях исходные данные используются без изменений. Лучший вариант с индивидуальным подбором поправки по каждой станции снизил RMSE источника с 1.6840 до 1.5749 °C на наборе из 125 станций (≈6.5%) и с 1.73 до 1.62 °C на наборе из 132

станций (≈6.3%). Конформная квантильная регрессия обеспечила фактическое покрытие 86% при целевом уровне 85%.

Процедура отбора поправки устроена следующим образом: для каждой станции сравниваются ошибки с поправкой и без на обучающих данных; поправка принимается только при устойчивом улучшении. Для станций с нестабильными данными поправка не применяется. Это предотвращает ухудшение точности там, где глобальная коррекция смещения неэффективна. Устойчивость результата дополнительно проверялась на нескольких последовательных годовых окнах: в каждом из них поправочный вариант не уступал исходному источнику.

Для оптимизации гиперпараметров XGBoost использована библиотека Optuna с методом байесовской оптимизации TPE. Оптимизировались: глубина деревьев, темп обучения, доля признаков и объектов в каждом дереве, сила L1- и L2-регуляризации. Поиск проводился по 100 итерациям с ранней остановкой, оценка — по RMSE на валидационном периоде 2020–2021 гг. Финальная модель обучалась на объединённой выборке 2013–2021 гг. с оптимальными гиперпараметрами.

Сводная таблица результатов содержит показатели RMSE, MAE и R^2 для шести вариантов: базовый ERA5 (точка отсчёта), XGBoost без дополнительных признаков, XGBoost с данными за предыдущие дни, финальная саратовская модель, результаты для Волгоградской области и результаты коррекции открытого источника. Финальная модель даёт снижение RMSE и MAE на 19.6% относительно ERA5 при росте R^2 с 0.987 до 0.993.

Дополнительно проверено, как влияет отсутствие спутниковых данных на точность прогноза. При разбивке тестовой выборки на три группы — строки с полными, частичными и отсутствующими данными MODIS — ошибка закономерно растёт при отсутствии спутниковых данных, однако во всех трёх группах модель точнее прямого использования ERA5.

ЗАКЛЮЧЕНИЕ

В работе разработана и проверена схема прогнозирования среднесуточной температуры воздуха, объединяющая данные реанализа ERA5, спутниковые измерения MODIS и наземные наблюдения Росгидромета за 2013–2023 годы. Отдельно рассмотрены перенос модели на Волгоградскую область и согласование открытого источника данных с официальными наблюдениями.

Первый результат получен на данных по Саратовской области. Итоговая модель снизила RMSE с 0.9806 до 0.7882 °C (≈19.6%) и MAE с 0.7437 до 0.5978 °C относительно прямого использования ERA5. Основной прирост дали лаговые признаки и координатное описание станций. Проверка важности признаков подтвердила, что модель не сводится к подстановке среднего по станции.

Второй результат связан с переносом на Волгоградскую область. Лучшее качество даёт обучение на целевом регионе: RMSE = 0.8328 °C и MAE = 0.6298 °C. Перенос без адаптации уступает дообучению и обучению с нуля. Следовательно, переносится не готовая модель, а выбранная схема признаков и порядок обучения.

Третий результат относится к калибровке RP5-подобного источника. Коррекция смещения с индивидуальным подбором поправки по каждой станции снизила RMSE источника с 1.6840 до 1.5749 °C на 125 станциях (≈6.5%). Конформная квантильная регрессия даёт фактическое покрытие около целевого уровня 0.85 без заметного роста средней ширины интервала.

Практическая ценность работы состоит в том, что предложенная методика полностью задокументирована: разбиения данных фиксированы, метрики сохраняются по каждому запуску, результаты проверяются в разрезе станций, сезонов и подвыборок. Схема может быть применена к другим регионам России при наличии архива ERA5, данных MODIS и наблюдений хотя бы нескольких станций Росгидромета. В качестве направлений дальнейшей работы представляются перспективными: улучшение обработки

пропусков MODIS с применением интерполяционных методов, добавление признаков рельефа местности и расширение проверки на другие регионы европейской части России.

Проведённая работа показала, что точность восстановления температуры воздуха по сеточным данным существенно зависит от набора признаков: добавление значений за предыдущие дни дало наибольший прирост точности, тогда как усложнение самого алгоритма обучения заметного эффекта не принесло. Полученный результат согласуется с выводами, сделанными для аналогичных задач в других регионах [3, 4]: именно признаковое описание объекта наблюдений, а не архитектура модели определяет итоговое качество прогноза на метеорологических данных.

Основные источники информации:

1. Hersbach H. et al. The ERA5 global reanalysis // Quarterly J. Royal Meteor. Soc. — 2020. — Vol. 146. — P. 1999–2049.
2. Mo Y. et al. Gap-filling methods for all-weather MODIS near-surface air temperature // Remote Sensing of Environment. — 2023. — Vol. 296.
3. Xin Y. et al. Mapping air temperature in China from time-normalized MODIS LST via stacking ensemble models // Geo-spatial Inf. Sci. — 2025.
4. Oduro C. et al. Leveraging machine learning for near-surface air temperature prediction in Ghana // J. African Earth Sci. — 2026. — Vol. 233.
5. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // KDD. — 2016. — P. 785–794.
6. Grinsztajn L. et al. Why do tree-based models still outperform deep learning on tabular data? // NeurIPS. — 2022. — Vol. 35. — P. 507–520.
7. Romano Y. et al. Conformalized Quantile Regression // NeurIPS. — 2019. — Vol. 32.