УДК 551.571, 004.852 DOI: 10.29006/1564-2291.JOR-2025.53(2).7

ВОССТАНОВЛЕНИЕ ПРИПОВЕРХНОСТНОЙ ВЛАЖНОСТИ АТМОСФЕРЫ НАД ОКЕАНОМ ПО ДАННЫМ СОПУТСТВУЮЩИХ МЕТЕОРОЛОГИЧЕСКИХ ИЗМЕРЕНИЙ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

С. А. Вострикова[1], М. А. Криницкий[1,2], С. К. Гулев[2], М. П. Александрова[2]

[1] Московский физико-технический институт,
Россия, 117303, Долгопрудный, Институтский переулок, д. 9,
e-mail: vostrikova.sa@phystech.edu;
[2] Институт океанологии им. П. П. Ширшова РАН,
Россия, 117997, Москва, Нахимовский проспект, д. 36,

Влажность воздуха в приповерхностном слое атмосферы над океаном является ключевым климатическим параметром, влияющим на процессы переноса влаги и тепла между океаном и атмосферой, а также на динамику атмосферных процессов в целом.

Анализ метеорологических данных, собранных в течение XX века, показывает разреженность рядов измерений влажности в пространстве и времени. Международный массив данных о характеристиках океана и атмосферы (ICOADS) указывает на недостаточную плотность измерений в начале XX века по сравнению с более поздними периодами, что создает сложности для адекватного анализа климатических тенденций относительной влажности.

Представленные в литературе методы восстановления временных рядов влажности зачастую демонстрируют ограниченную точность, основываясь преимущественно на статистических и эвристических подходах. Наша работа направлена на повышение качества решения этой задачи за счет применения методов машинного обучения.

В настоящей статье решена задача в формулировке аппроксимации моментальных значений относительной влажности по данным сопутствующих измерений атмосферного давления, температуры воздуха, скорости и направления ветра, температуры поверхности океана, а также наблюдений количества и типов облачности на трех ярусах. Кроме этого, в составе сопутствующих переменных используется код погоды по стандарту ВМО и расчетная высота солнца. В исследовании использованы модели машинного обучения следующих типов: линейная регрессия, дерево решений, случайный лес, градиентный бустинг и полносвязная искусственная нейронная сеть. Для повышения территориальной и временной специфичности разрабатываемых моделей мы провели исследование для каждой ячейки размером 2° по широте и долготе (географическая трапеция) и каждого сезона по отдельности. На основе полученных результатов были построены карты пространственного распределения ошибок моделей, которые позволили выявить регионы с высокой и низкой точностью аппроксимации влажности.

Исследование подтвердило эффективность методов машинного обучения для восстановления климатических рядов, определило наиболее подходящие модели для этой задачи и обозначило перспективные направления для дальнейшей работы.

Ключевые слова: относительная влажность, реконструкция климатических рядов, машинное обучение, восстановление регрессии, линейная регрессия, дерево решений, случайный лес, градиентный бустинг, искусственная нейронная сеть

1. Введение

Влажность воздуха над океаном в приповерхностном слое атмосферы – важный климатический параметр, оказывающий значительное влияние на динамику атмосферы и обменные процессы между океаном и атмосферой.

Ряды измерений относительной влажности *in situ* за XX столетие характерны неравномерным наполнением (Александрова, Гулев, 2023). Данные международного массива измерений характеристик океана и атмосферы (International Comprehensive Ocean-Atmosphere Data Set, ICOADS (Freeman et al., 2017)) показывают, что начало XX века недостаточно обеспечено измерениями по сравнению с периодами, начиная с середины столетия (Александрова, Гулев, 2023). Такая неравномерность измерений препятствует адекватному анализу климатических тенденций относительной влажности как важнейшей характеристики атмосферы. Восстановление рядов измерений относительной влажности может стать ключевым фактором в понимании климатических тенденций прошлого и современного климата.

В литературе существуют работы, демонстрирующие возможность аппроксимации относительной влажности по сопутствующим измерениям, ряды которых более наполнены по сравнению с рядами относительной влажности (Luminda Niroshana Gunawardhana et al., 2017; Jiajun Guo et al., 2023).

В одной из работ (Luminda Niroshana Gunawardhana et al., 2017) описывается подход к оценке суточной относительной влажности воздуха, основанный на использовании статистически скорректированных данных о минимальной температуре из глобальной климатической модели (GCM). В этом исследовании авторы устраняют систематические погрешности в модельных данных, сопоставляя их с реальными измерениями, чтобы получить скорректированную минимальную температуру. Затем, имея многолетние наблюдения о фактической влажности и локальной минимальной температуре, строят для каждого месяца эмпирические (линейные или квадратичные) зависимости между этими двумя показателями. Используя полученные формулы и скорректированные данные о температуре, вычисляют влажность для тех же суточных временных интервалов.

В другом исследовании (Jiajun Guo et al., 2023) предложен гибридный метод аппроксимации относительной влажности на основе метеорологических данных. В его основе лежат три составляющие: SARIMA (Seasonal Autoregressive Integrated Moving Average), используемая для выявления и моделирования сезонных колебаний относительной влажности; коинтеграционный анализ (Engle-Granger, EG), доказывающий наличие долгосрочной связи между влажностью и используемыми в работе метеорологическими параметрами; модель коррекции ошибок (Error Correction Model, ECM), анализирующая краткосрочные отклонения относительной влажности от долгосрочной траектории и определяющая факторы, которые в большей степени влияют на эти колебания.

В упомянутых работах применяются классические статистические методы или эвристические алгоритмы (Luminda Niroshana Gunawardhana et al., 2017; Jiajun Guo

et al., 2023). Для повышения точности восстановления относительной влажности в настоящем исследовании мы предлагаем новый подход к решению этой задачи, основанный на применении методов машинного обучения.

Целью этой работы является разработка и оценка моделей машинного обучения, способных аппроксимировать значения относительной влажности на основе доступных метеорологических данных. В терминах машинного обучения эта задача представляет собой задачу восстановления регрессии, в которой целевой переменной является относительная влажность, а признаками – сопутствующие метеорологические наблюдения и измерения.

На первом этапе мы сформировали набор данных DISO3 (**D**ataset of **I**n **S**itu **O**bservations **O**ver the **O**cean) на основе данных ICOADS, включающий попутные метеорологические измерения и наблюдения, а также соответствующие им расчетные значения относительной влажности. Эта база данных стала основой для обучения и оценки качества алгоритмов, основанных на подходе машинного обучения.

На втором этапе мы применили модели машинного обучения, для каждой из которых мы провели оптимизацию гиперпараметров – параметров, задающих конфигурацию модели и процесса ее обучения, значения которых выбираются до начала обучения и не обновляются в ходе оптимизации модели по данным. Подбор гиперпараметров позволил улучшить качество аппроксимации и обеспечить высокую обобщающую способность моделей.

На заключительном этапе мы исследовали обученные модели на предмет пространственного распределения точности. Анализ карт ошибок в сравнении с картами климатических характеристик (средней относительной влажности и ее изменчивости) позволит оценить связь отклонений моделей с экстремальными значениями влажности или ее вариабельностью.

2. Данные и методы

2.1. Данные

В настоящем исследовании использовался набор данных DISO3 – авторская база данных, сформированная на основе тщательно отобранных и предобработанных данных из массива ICOADS (Freeman et al., 2017), который представляет собой самый крупный и полный массив метеорологических и океанографических данных, собранных за длительный период времени. ICOADS содержит данные о морской поверхности, сопутствующие метеорологические наблюдения и измерения за период с 1662 г. по настоящее время. Пространственная плотность наблюдений и измерений базы данных ICOADS за период с 1950 по 2017 гг. показана на рисунке 1.

Историческая часть набора ICOADS включает записи, сделанные вручную в корабельных журналах, начиная с XVII века. С развитием технологий к ICOADS стали добавляться данные, собранные с буев, судов, береговых станций и

океанографических датчиков. Это позволило существенно повысить точность наблюдений, расширить их географическое покрытие и повысить пространственную плотность измерений.

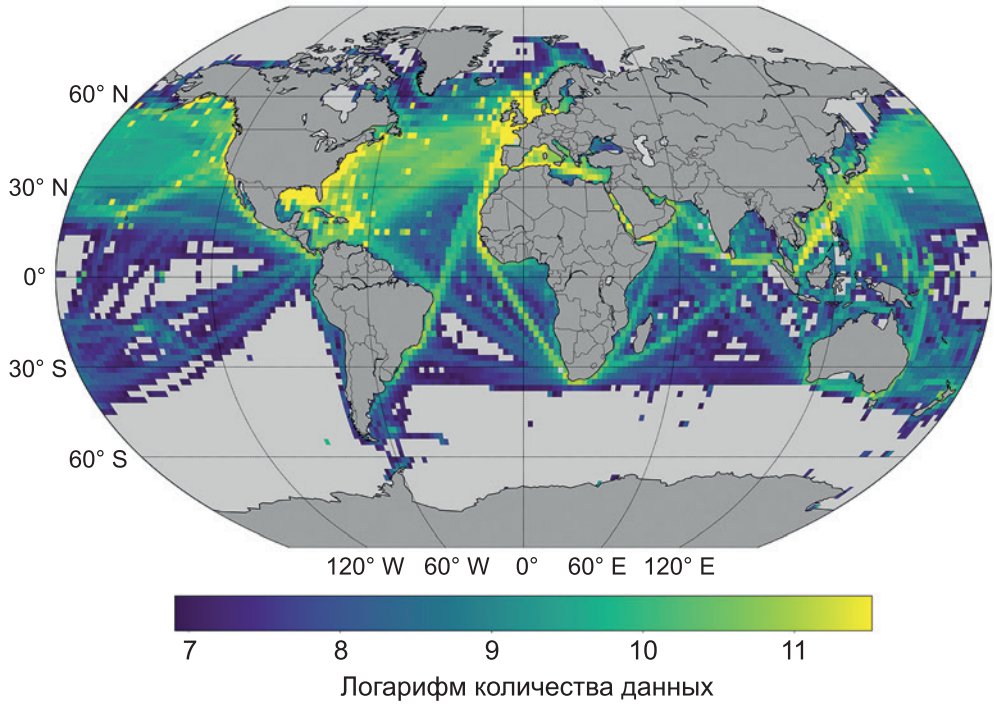

Рис. 1 – Пространственная плотность наблюдений и измерений в базе данных ICOADS (Freeman et al., 2017) за период с 1950 по 2017 гг., в логарифмической шкале. Дискретизация по пространству – 2×2°

ICOADS охватывает практически всю поверхность Мирового океана, однако плотность варьируется в зависимости от региона и времени. В периоды до XX века наблюдения были ограничены активностью судоходства и чаще всего проводились в популярных морских маршрутах. Начиная с середины XX века, наблюдения становятся более систематичными, благодаря внедрению автоматических систем наблюдений и международному сотрудничеству в области климатического мониторинга.

В DISO3 включены только тщательно отфильтрованные и скорректированные данные ICOADS, что позволяет избежать ошибок, связанных с некачественными записями, характерными для исторических данных (Александрова, Гулев, 2023).

Фильтрация заключалась в удалении записей с пропущенными значениями основных параметров, таких как дата, координаты, температура воздуха и атмосферное давление. После фильтрации для каждого наблюдения были рассчитаны производные параметры. Относительная влажность определялась на основе температуры воздуха, точки росы и давления с использованием приближенного эмпирического соотношения. Высота солнца над горизонтом рассчитывалась по координатам, дате и времени суток на основе геометрической модели астрономического положения Солнца.

DISO3 охватывает период с 1880 по 2017 гг. При этом в настоящей работе для обучения и оценки качества моделей были использованы данные за период с 1950 по 2017 гг. (93 млн записей). На рисунке 2 показано общее годовое количество наблюдений в наборе данных DISO3.

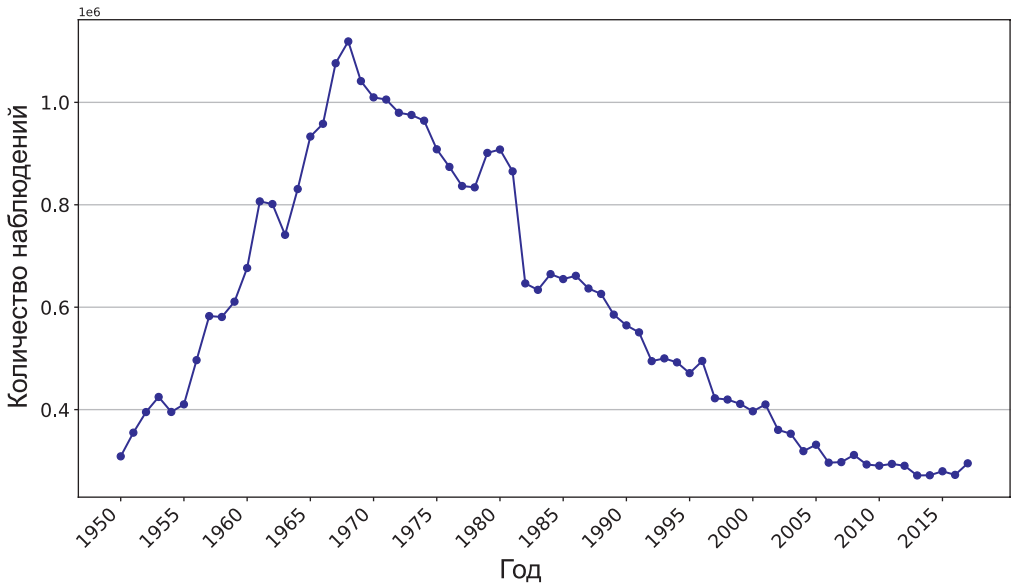

Рис. 2 – Общее годовое количество наблюдений в наборе данных DISO3

Набор данных DISO3 включает подмножество переменных, представленных в ICOADS. В таблице 1 мы приводим состав переменных, включенных в DISO3.

Табл. 1 – Состав переменных набора данных DISO3

Переменная	**Единицы измерения**	**Наименование**
Высота солнца над горизонтом	градусы	hsun
Атмосферное давление	гПа	slp
Температура атмосферы	°C	ta
Температура поверхности океана	°C	sst
Балл общей облачности	окты	icn
Балл облачности нижнего яруса	окты	icl
Тип облачности нижнего яруса	класс согласно классификации ICOADS [ref]	low
Тип облачности среднего яруса	класс согласно классификации ICOADS [ref]	mid
Тип облачности верхнего яруса	класс согласно классификации ICOADS [ref]	high
Код погоды	код согласно классификации ВМО [ref]	iww
Скорость ветра	м/с	ws
Направление ветра	градусы от направления на север	wd
Относительная влажность	доли единиц	rh

Для корректной работы алгоритмов машинного обучения, чувствительных к масштабу входных данных, все действительные признаки (числовые признаки с непрерывными значениями) были стандартизированы. Под стандартизацией понимается обезразмеривание признаков путем вычитания выборочного среднего и деления

на выборочное стандартное отклонение. Это позволяет привести данные к единому масштабу, где каждое значение отражает отклонение от среднего в единицах стандартного отклонения.

Формула стандартизации для значения признака X выглядит следующим образом:

$$X^* = \frac{X - \mu}{\sigma}, \quad (1)$$

где X^*– стандартизированное значение признака, X – исходное значение признака, μ – выборочное среднее признака, σ – выборочное стандартное отклонение признака X. В результате такого преобразования значения действительных признаков имеют среднее 0 и стандартное отклонение 1.

Также стандартизация данных позволила приблизить их распределение к нормальному и заполнить пропуски в данных значениями, сгенерированными из этого распределения. Мы заполняли пропуски сгенерированными значениями для действительных переменных (скорость и направление ветра). При этом пропущенные значения для категориальных переменных (балл общей облачности, балл облачности нижнего яруса, тип облачности нижнего, среднего, верхнего яруса и код погоды) удалены.

2.2. Методы

2.2.1. Учет пространственно-временной изменчивости приповерхностной влажности

Приповерхностная влажность атмосферы зависит от множества факторов, которые сильно различаются в разных частях океана (Shakespeare, Roderick, 2024). Кроме того, в ряде регионов Мирового океана относительная влажность приповерхностного слоя атмосферы характеризуется существенными сезонными вариациями.

Использование глобальных данных в рамках подхода обучения статистических моделей (моделей машинного обучения) зачастую приводит к усреднению, скрывающему региональные и сезонные особенности (Borchert, 2008; Abraham, Goldblatt, 2023). Поэтому для повышения пространственной и временной специфичности применен подход, основанный на разбиении всей водной поверхности на ячейки размером 2×2 градуса по широте и долготе. Это позволяет проводить региональные оценки относительной влажности. Пространственные ячейки формировались на основе регулярной широтно-долготной сетки с шагом 2°, где границы каждой ячейки определяются целыми значениями широты и долготы. Такой способ построения обеспечивает непрерывное и неперекрывающееся покрытие всей исследуемой поверхности океана.

Дальнейшее улучшение точности достигается за счет разделения данных по сезонам внутри каждой ячейки, что позволяет учесть сезонные колебания относительной влажности. Из рассмотрения были исключены регионы, в которых меньше 1000 наблюдений, поскольку такого количества данных недостаточно для надежного

обучения моделей машинного обучения. Кроме того, для ускорения обучения моделей в регионах с более, чем 100 000 данных, выборка была сокращена до 100 000 наблюдений. Распределение количества наблюдений в ячейках 2×2° в наборе данных DISO3 представлено на рисунке 3.

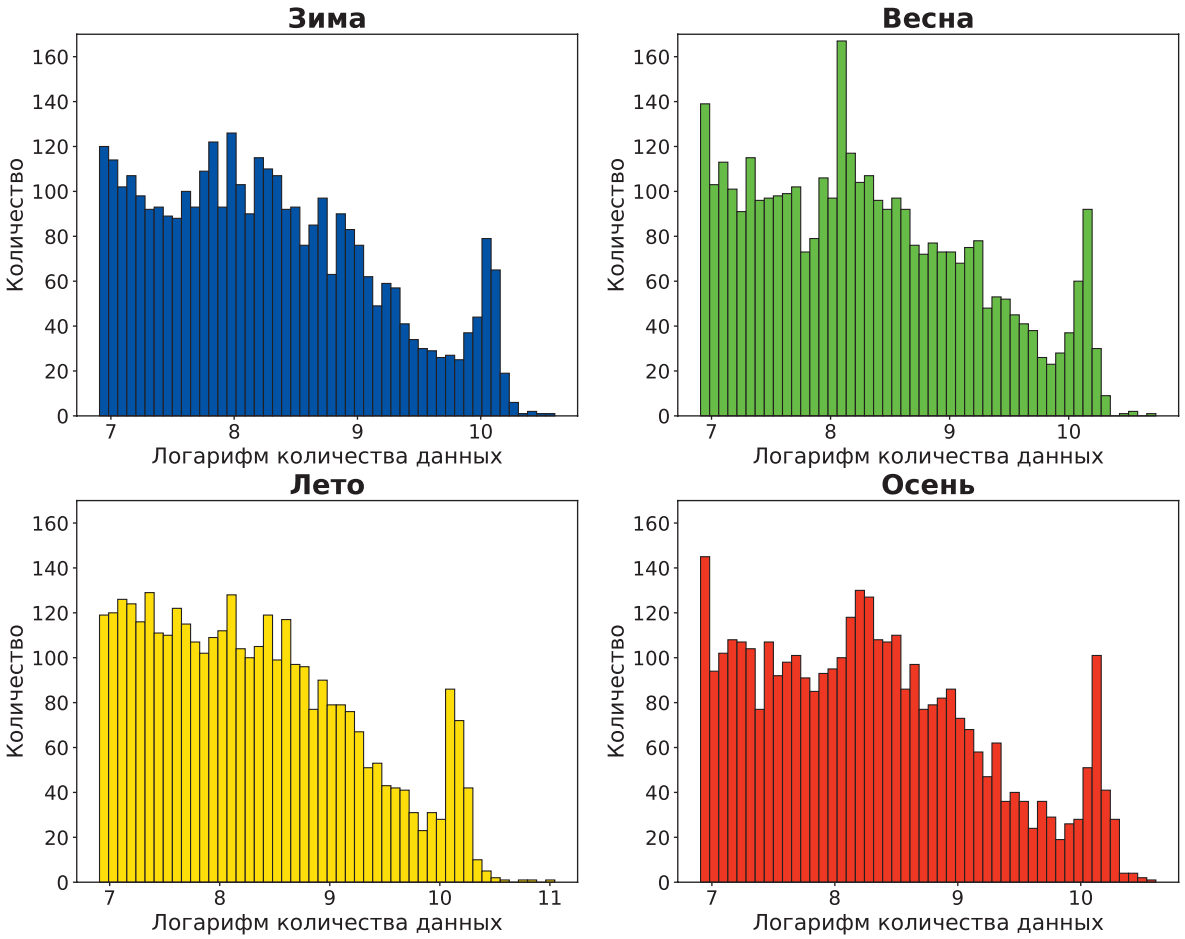

Рис. 3 – Распределение количества наблюдений в ячейках 2×2° в наборе данных DISO3

2.2.2. Подбор оптимальных гиперпараметров для моделей

Для того, чтобы минимизировать ошибку аппроксимации относительной влажности и повысить обобщающую способность алгоритмов машинного обучения, необходимо подобрать оптимальные гиперпараметры каждой модели. Для достижения этой цели была использована библиотека байесовской оптимизации Optuna (Akiba et al., 2019; Almarzooq, Waheed, 2024). В качестве алгоритма сэмплирования выбран TPESampler – алгоритм, основанный на построении вероятностных моделей распределения гиперпараметров.

Выбор TPESampler обусловлен его эффективностью в задачах с высокой размерностью пространства гиперпараметров (Parizy et al., 2023). Этот алгоритм рационально использует информацию из предыдущих итераций, что позволяет быстрее сходиться к оптимальным значениям, требуя меньшего числа оценок целевой функции.

Его байесовский подход позволяет эффективно исследовать пространство поиска, избегая ненужных вычислений в неперспективных областях.

2.2.3. Обучение моделей

Для исследования были выбраны пять моделей машинного обучения: линейная регрессия, дерево решений (Barry de Ville, 2013), случайный лес (Cutler et al., 2012) градиентный бустинг в реализации CatBoost (Hancock, Khoshgoftaar, 2020) и полносвязная искусственная нейронная сеть (Minsky, Papert, 1969). Обучение каждой модели выполнялось отдельно для каждой 2×2-градусной ячейки с учетом разделения данных по сезонам. Для обучения моделей использовалось 70 % данных, упорядоченных по времени, а оставшиеся 30 % были отложены для оценки качества, сохраняя таким образом хронологическую структуру.

Приведем краткое описание принципа работы алгоритмов машинного обучения, которые использовались в работе.

Линейная регрессия – это один из наиболее простых методов машинного обучения, предназначенный для аппроксимации непрерывной переменной на основе независимых переменных. Метод основан на предположении о линейной зависимости между целевой переменной и признаками. Модель оптимизируется методом наименьших квадратов, который минимизирует сумму квадратов отклонений между аппроксимированными и фактическими значениями целевой переменной.

Главное преимущество линейной регрессии – ее вычислительная эффективность, которая позволяет обрабатывать большие объемы данных за приемлемое время. Однако из-за предположения о линейности эта модель не способна улавливать нелинейные зависимости между признаками и целевой переменной, что приводит к снижению точности в тех случаях, когда реальные зависимости более сложные.

Дерево решений – это алгоритм машинного обучения, который строит иерархическую структуру (дерево) для последовательного разделения данных на подмножества. Каждое ветвление дерева решений соответствует правилу, основанному на значении одного из атрибутов данных. Узлы дерева выполняют проверку значения конкретного атрибута, а листья представляют собой итоговые регрессионные значения. Процесс построения дерева начинается с корневого узла, который содержит всю исходную выборку данных. Алгоритм итеративно выбирает наилучший атрибут для разделения данных на основе определенного критерия качества. Для задач регрессии этим критерием обычно является среднеквадратичная ошибка MSE, т. е. средний квадрат отклонений аппроксимированных значений от фактических в каждом подмножестве. Оптимальный атрибут – тот, который обеспечивает максимальное уменьшение MSE после разбиения. После выбора атрибута данные разделяются на подмножества, соответствующие различным значениям атрибута. Этот процесс повторяется для каждого подмножества, создавая новые узлы и ветви дерева и продолжается до тех пор, пока не будет достигнуто заданное условие остановки. Этим условием может быть, например, достижение максимальной глубины дерева или минимального числа

элементов в узле. Процесс обучения деревьев решений относительно быстрый, что делает их эффективными при работе с большим набором данных. Более того, деревья решений эффективно справляются с нелинейными взаимосвязями между признаками и целевой переменной, в отличие от моделей, предполагающих линейность. Однако у этой модели есть существенный недостаток – склонность к переобучению: модель слишком хорошо запоминает обучающие данные, из-за чего демонстрирует низкую обобщающую способность. Использование деревьев решений в ансамблях решает проблему переобучения.

Случайный лес – это алгоритм ансамблевого обучения, который усредняет предсказания множества отдельных решающих деревьев. В отличие от использования одного сложного решающего дерева, подверженного переобучению, случайный лес минимизирует этот риск за счет создания множества более простых, но разнообразных деревьев и усреднения их результатов. На первом этапе строится множество решающих деревьев. Во-первых, каждое дерево создается на случайной подвыборке исходного набора данных. Во-вторых, при построении узлов каждого дерева случайный лес рассматривает не все признаки, а только случайное подмножество из них. На втором этапе каждое дерево в случайном лесу независимо формирует свою оценку целевой переменной. Итоговый результат получается путем усреднения этих оценок. Случайный лес обладает рядом преимуществ перед другими алгоритмами машинного обучения. Высокая точность аппроксимации обусловлена ансамблевым подходом и низкой корреляцией между составляющими его деревьями. Кроме того, случайный лес демонстрирует высокую устойчивость к переобучению благодаря случайности в процессе построения деревьев. Это делает его эффективным для работы даже со сложными, зашумленными данными, содержащими выбросы. Также алгоритм поддерживает параллельные вычисления, что позволяет существенно ускорить обучение на больших объемах данных.

CatBoost – это алгоритм ансамблевого обучения, представляющий собой усовершенствованную реализацию градиентного бустинга. Его ключевое преимущество – эффективный способ обработки категориальных признаков, который значительно повышает точность аппроксимации по сравнению с традиционными методами. Вместо стандартных подходов, CatBoost использует собственный алгоритм, основанный на построении вероятностных моделей для каждого категориального признака, учитывая его влияние на целевую переменную. Алгоритм строит последовательность деревьев, каждое из которых исправляет ошибки предыдущих, при этом порядок данных для обучения случайным образом меняется перед построением каждого дерева. Такой подход уменьшает влияние порядка данных на результат, делая модель более устойчивой к шуму и выбросам. Главное преимущество CatBoost в том, что он не требует предварительной обработки категориальных переменных, что упрощает процесс подготовки данных. Кроме того, CatBoost устойчив к переобучению благодаря использованию упорядоченного бустинга. Также он эффективно обрабатывает большие наборы данных и поддерживает параллельные вычисления, что существенно сокращает время обучения модели.

Полносвязная нейронная сеть – это тип искусственной нейронной сети, в которой каждый нейрон одного слоя связан со всеми нейронами следующего слоя. Алгоритм можно разделить на два основных этапа: прямой проход и обратный проход. Прямой проход начинается с подачи входных данных, представляющих собой вектор признаков, в первый слой, называемый входным. Каждый нейрон входного слоя передает значение соответствующего признака на следующий слой. В последующих слоях каждый нейрон выполняет взвешенное суммирование значений, полученных от нейронов предыдущего слоя, добавляет к результату смещение и пропускает полученную сумму через активационную функцию. Активационная функция вносит нелинейность в модель, позволяя нейронной сети аппроксимировать сложные нелинейные зависимости между признаками и целевой переменной. В данной работе была выбрана активационная функция ReLU, исходя из ее вычислительной эффективности, устойчивости к затуханию градиента и способности повышать обобщающую способность модели за счет разреженности активаций. После прохождения сигнала через все слои нейронной сети на выходе получаем значение целевой переменной. Обратный проход служит для коррекции весов и смещений, минимизируя ошибку аппроксимации. Для этого используется алгоритм обратного распространения ошибки. Суть алгоритма заключается в вычислении градиента функции ошибки по весам и смещениям каждого нейрона. Градиент показывает, как изменение весов и смещений влияет на величину ошибки. Затем, используя алгоритм оптимизации (в данной работе выбран Adam, ввиду его высокой эффективности и устойчивости), веса и смещения корректируются в направлении, уменьшающем ошибку. Процесс обратного прохода включает вычисление частных производных функции ошибки по весам и смещениям, распространяя ошибку от выходного слоя к входному. Этот процесс повторяется итеративно на протяжении всего обучения, пока не будет выполнено заданное количество итераций. В процессе обучения нейронная сеть находит оптимальные веса и смещения, которые позволяют ей наиболее точно аппроксимировать целевую переменную на основе входных данных. Преимущество полносвязных нейронных сетей в том, что они способны моделировать сложные, нелинейные зависимости в данных с высокой точностью. Кроме того, ее гибкость позволяет легко масштабировать модель, изменяя количество нейронов и слоев для адаптации к различным объемам данных и сложности задачи. Также простая интеграция методов регуляризации, например, dropout, позволяет эффективно бороться с переобучением и улучшать обобщающую способность модели.

2.2.4. Оценка качества моделей

Эффективность алгоритмов мы оценивали на основе нескольких критериев, включающих:

• количественную оценку: определено качество каждой модели на отложенной для теста выборке с использованием метрики качества;

• оценку неопределенностей: проведена оценка неопределенности меры качества и целевой переменной, что позволило сделать выводы о надежности полученных результатов;

• сравнение с тривиальной моделью: качество аппроксимации для каждой модели сравнено с качеством для тривиальной модели – базового алгоритма, который не использует входные признаки и аппроксимирует влажность, как среднее значение величины по региону. Это показало, насколько эффективно работает модель машинного обучения по сравнению с простым усреднением.

На основании этих оценок выбран алгоритм с наилучшими показателями для дальнейшего его применения.

Качество аппроксимации оценивалось на отложенной для теста выборке. Метрикой качества была принята среднеквадратичная ошибка RMSE (2). Эта величина дает интерпретируемое и наглядное представление о том, насколько хорошо модель аппроксимирует данные, поскольку выражается в тех же единицах измерения, что и целевая переменная.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}, \qquad (2)$$

где N – количество наблюдений в выборке, y_i – фактическое значение целевой переменной для i-го наблюдения, $\hat{y}_i$ – полученное моделью значение целевой переменной для i-го наблюдения.

2.2.5. Анализ ошибок и сравнение моделей

По полученным результатам были построены пространственные карты ошибок для каждого сезона. Они отображают географическое распределение RMSE моделей.

Эти карты сопоставлены с климатическими картами относительной влажности (показывающими среднее значение по региону) и картами ее изменения. Сравнение карт позволило выявить зависимости между величиной ошибок аппроксимации и характеристиками климата. В частности, такой анализ помог понять, связаны ли большие ошибки с экстремальными значениями относительной влажности или с ее высокой вариабельностью.

Также были построены гистограммы распределения ошибок, которые позволили оценить стабильность моделей по разбросу значений RMSE.

Для оценки эффективности моделей машинного обучения их результаты были сопоставлены с тривиальной моделью, представляющей собой простейший подход к восстановлению относительной влажности: она аппроксимирует ее как среднее значение величины по региону. Сравнение проводилось посредством вычисления относительных ошибок, что позволило оценить, насколько значимо улучшение качества аппроксимации, достигнутого с помощью моделей машинного обучения, а также определить, оправдывает ли их сложность и вычислительные затраты повышение точности.

2.2.6. Оценка неопределенностей меры качества и целевой переменной

Меры качества дают усредненное представление о точности моделей, позволяя оценить, насколько хорошо они в целом аппроксимируют целевую переменную. Однако такие меры не отражают степень надежности полученных результатов. Для более точной оценки моделей необходимо учитывать неопределенности как меры качества, так и целевой переменной. Для этого в данной работе был использован подход бутстреп (Efron, 1979).

Бутстреп – это метод сэмплирования из выборки с возвращением, позволяющий оценивать неопределенности на основе множества подвыборок.

Процедура bootstrap начинается с определения размера bootstrap-выборки, которая, как правило, равна размеру исходной выборки. Затем определенное количество раз производится случайный отбор элементов из исходной выборки с возвращением. Это означает, что после выбора элемента он возвращается обратно в исходную выборку, и, следовательно, один и тот же элемент может быть выбран несколько раз в одной bootstrap-выборке, а некоторые элементы исходной выборки могут вообще не попасть в конкретную выборку (OOB-выборка). В данной работе в каждой генерируемой bootstrap-выборке выдерживалась хронологическая последовательность данных, что обеспечивало сохранение временной структуры исходных данных.

Для оценки неопределенности качества модели на каждой bootstrap-выборке обучается отдельная модель, RMSE которой оценивается на ее OOB-выборке. Итоговое качество модели определяется как среднее значение оценок, полученных на всех OOB-выборках, что позволяет учесть как случайный шум в данных (алеаторическая неопределенность), так и ограничения модели (эпистемическая неопределенность).

Для оценки неопределенности целевой переменной на каждой bootstrap-выборке обучается отдельная модель и сохраняются значения ее параметров. К каждому новому объекту применяются все обученные модели. Неопределенность целевой переменной оценивается как выборочное стандартное отклонение полученного множества оценок. Это стандартное отклонение отражает эпистемическую неопределенность модели.

3. Результаты

3.1. Оценка RMSE для моделей

Табл. 2 – Среднее значение RMSE для моделей.
Лучшая модель выделена жирным шрифтом

Модель	**Среднее (по океану) RMSE**			
	Зима	**Весна**	**Лето**	**Осень**
Тривиальная модель	0.1050	0.1018	0.0890	0.1003
Линейная регрессия	0.1017	0.0974	0.0844	0.0955
Дерево решений	0.1021	0.0979	0.0849	0.0962
Случайный лес	0.0997	0.0952	0.0827	0.0935
Полносвязная нейросеть	0.0996	0.0951	0.0824	0.0932
Catboost	**0.0989**	**0.0943**	**0.0818**	**0.0926**

В таблице показано среднее качество RMSE для моделей, обученных с оптимальными гиперпараметрами и протестированных на отложенных выборках. По значениям RMSE можно сделать вывод, что Catboost – лучшая модель по среднему качеству на тестовой выборке.

3.2. Оценка неопределенности меры качества и целевой переменной

Табл. 3 – Среднее (по океану) значение RMSE по bootstrap-выборкам.
Лучшая модель выделена жирным шрифтом

Модель	**Среднее (по океану) RMSE по bootstrap-выборкам (в долях)**			
	Зима	**Весна**	**Лето**	**Осень**
Тривиальная модель	0.10528	0.10210	0.08981	0.10137
Линейная регрессия	0.09614	0.09210	0.08028	0.09036
Дерево решений	0.09660	0.09239	0.08073	0.09096
Случайный лес	**0.08560**	**0.08075**	**0.07127**	**0.07944**
Полносвязная нейросеть	0.09651	0.09210	0.08097	0.09040
Catboost	0.09556	0.09122	0.07894	0.09013

Табл. 4 – Неопределенность меры качества в мере среднего по всему Мировому океану стандартного отклонения. Приведено в долях единицы.
Лучший показатель за каждый сезон выделен жирным шрифтом

Модель	**Мера неопределенности**			
	Зима	**Весна**	**Лето**	**Осень**
Тривиальная модель	**0.00018**	**0.00017**	**0.00017**	**0.00018**
Линейная регрессия	0.00030	0.00029	0.00030	0.00029
Дерево решений	0.00056	0.00059	0.00054	0.00060
Случайный лес	0.00066	0.00067	0.00061	0.00067
Полносвязная нейросеть	0.00092	0.00093	0.00086	0.00088
Catboost	0.00084	0.00086	0.00072	0.00076

В таблицах представлены значения среднего по bootstrap-выборкам качества моделей и среднее стандартное отклонение RMSE. Случайный лес – наилучшая модель по среднему качеству на bootstrap-выборках. При этом неопределенность меры качества низкая для всех моделей.

Табл. 5 – Неопределенность целевой переменной (относительной влажности) в мере среднего по всему океану стандартного отклонения. Приведено в долях единицы.
Лучший показатель за каждый сезон выделен жирным шрифтом

Модель	**Мера неопределенности**			
	Зима	**Весна**	**Лето**	**Осень**
Тривиальная модель	**0.00462**	**0.00444**	**0.00415**	**0.00436**
Линейная регрессия	0.00843	0.00829	0.00743	0.00800
Дерево решений	0.01808	0.01847	0.01585	0.01867
Случайный лес	0.01287	0.01312	0.01114	0.01305
Полносвязная нейросеть	0.01434	0.01456	0.01245	0.01411
Catboost	0.01080	0.01078	0.00906	0.01045

В данной таблице приведены значения среднего стандартного отклонения целевой переменной. Тривиальная модель, линейная регрессия и Catboost показали наименьшую неопределенность целевой переменной.

3.3. Оценка относительных ошибок

Табл. 6 – Относительные ошибки моделей.
Лучший показатель за каждый сезон выделен жирным шрифтом

Модель	**Относительная ошибка (в %)**			
	Зима	**Весна**	**Лето**	**Осень**
Линейная регрессия	2.95	4.47	5.15	4.83
Дерево решений	2.78	3.87	4.64	4.18
Случайный лес	5.05	6.49	7.00	6.70
Полносвязная нейросеть	**5.81**	**7.39**	**8.03**	**7.59**
Catboost	5.27	6.75	7.48	7.08

В таблице 6 приведены значения относительных ошибок моделей. Полносвязная нейросеть, Catboost и случайный лес достигли существенно более высокого уровня качества по сравнению с тривиальной моделью.

3.4. Анализ пространственных карт и гистограмм

В этом разделе помещаем результаты оценки ошибок всех моделей, используемых в настоящем исследовании. Мы приводим эти результаты в форме агрегированных ошибок (сезонных гистограмм ошибок в целом по всему Мировому океану) и в форме карт ошибок. Гистограммы ошибок приведены на рисунках 8 – для тривиальной модели, 10 – для модели линейной регрессии, 12 – для модели дерева решений, 14 – для модели случайных лесов, 16 – для модели градиентного бустинга в реализации CatBoost и 18 – для модели искусственной нейронной сети. Карты ошибок приведены на рисунках 9 – для тривиальной модели, 11 – для модели линейной регрессии, 13 – для модели дерева решений, 15 – для модели случайных лесов, 17 – для модели CatBoost и 19 – для модели искусственной нейронной сети. Для сравнения мы также приводим климатические характеристики влажности по данным DISO3 в целом по океану и в пространственном распределении. Так, на рисунке 4 мы приводим гистограмму климатических средних относительной влажности (в долях единицы); на рисунке 5 отображено пространственное распределение климатических средних относительной влажности атмосферы над Мировым океаном по данным DISO3. На рисунках 6 и 7 мы приводим гистограммы и карты дисперсии относительной влажности, характеризующие вариативность относительной влажности в ячейке 2×2°.

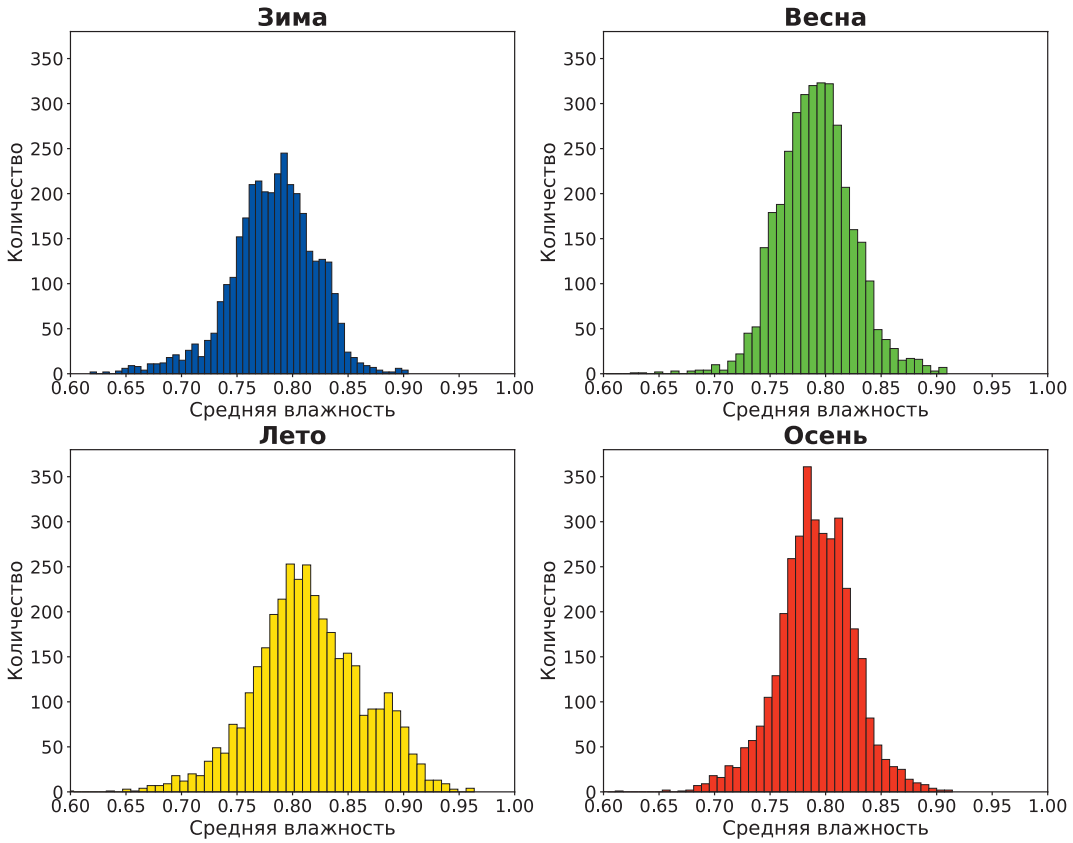

Рис. 4 – Гистограммы климатических средних относительной влажности по сезонам, по данным DISO3

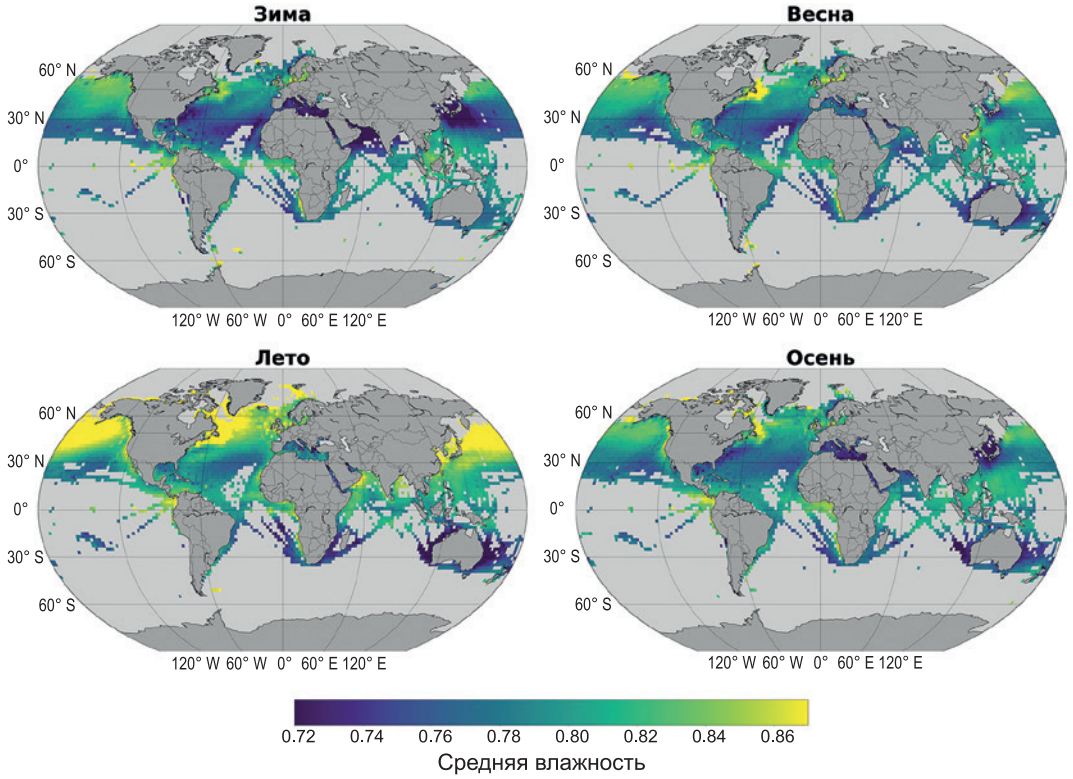

Рис. 5 – Карты климатических средних относительной влажности по сезонам, по данным DISO3

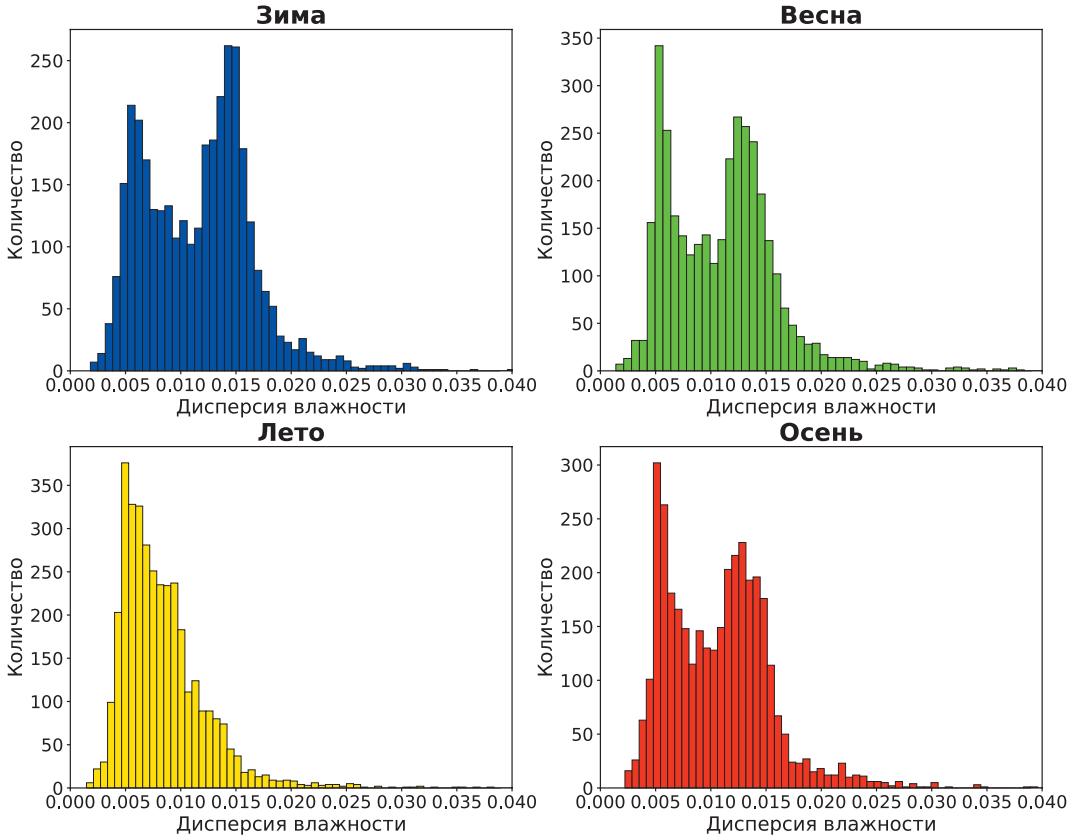

Рис. 6 – Гистограммы разброса значений относительной влажности по сезонам, по данным DISO3

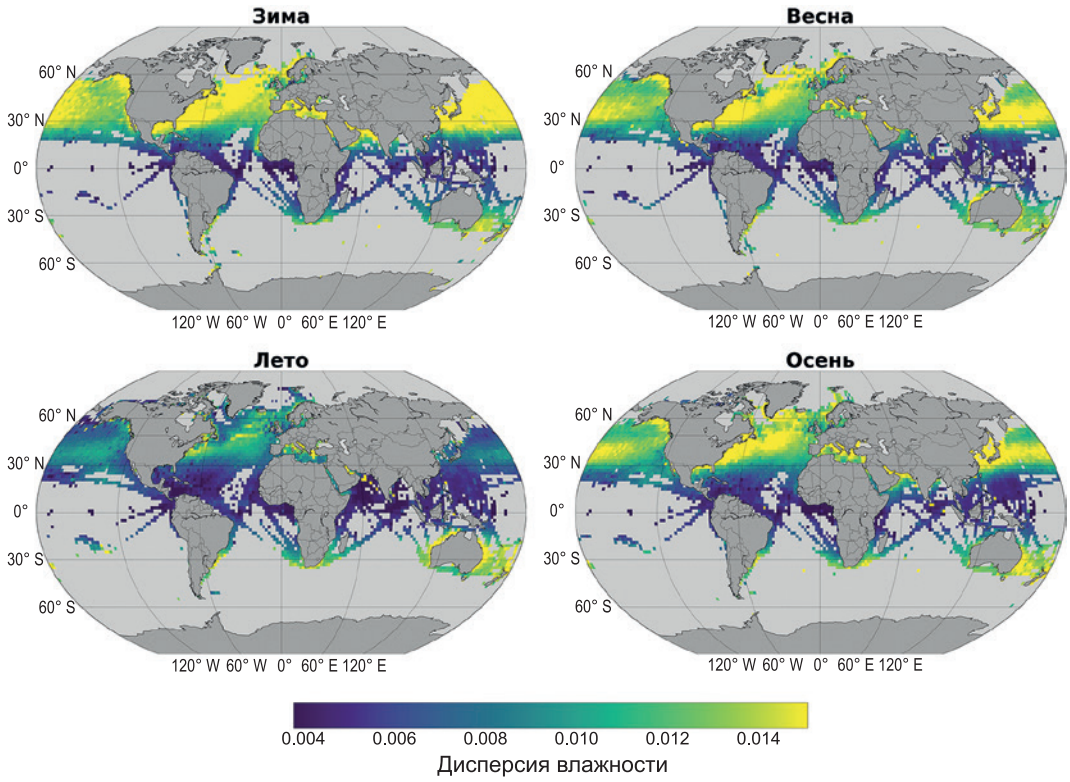

Рис. 7 – Карты разброса значений относительной влажности по сезонам

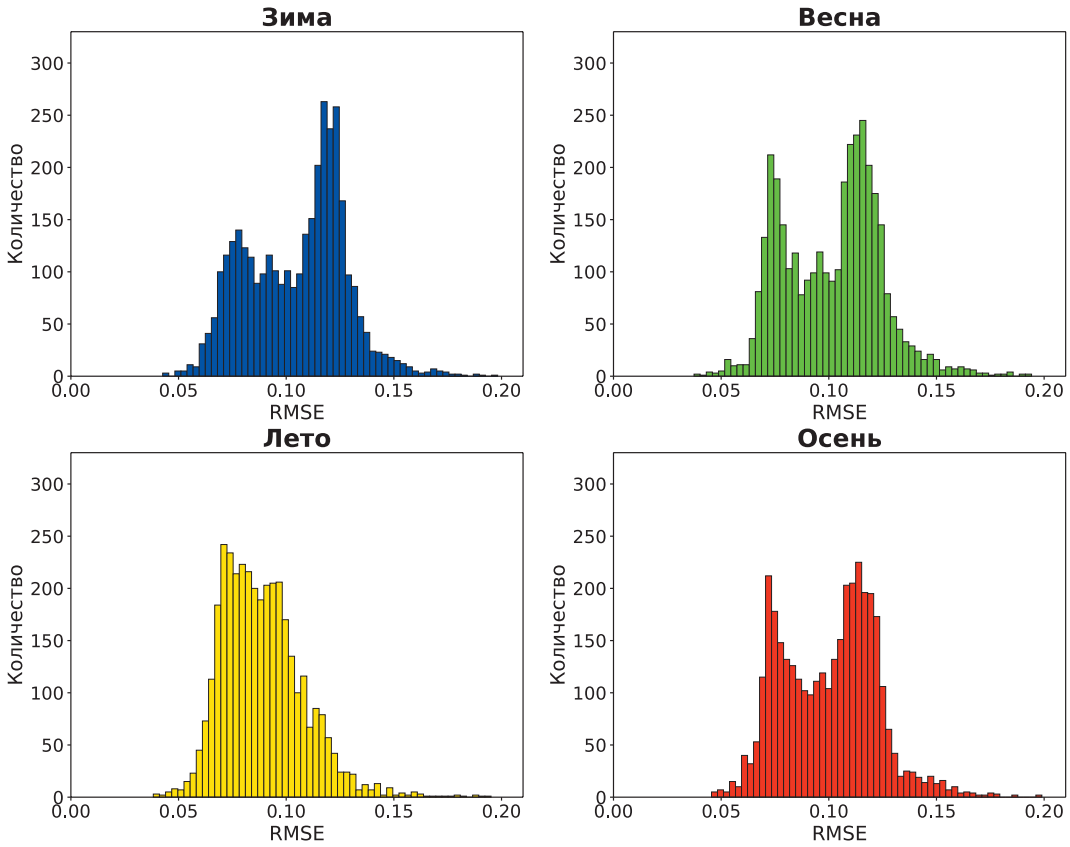

Рис. 8 – Гистограммы ошибок тривиальной модели

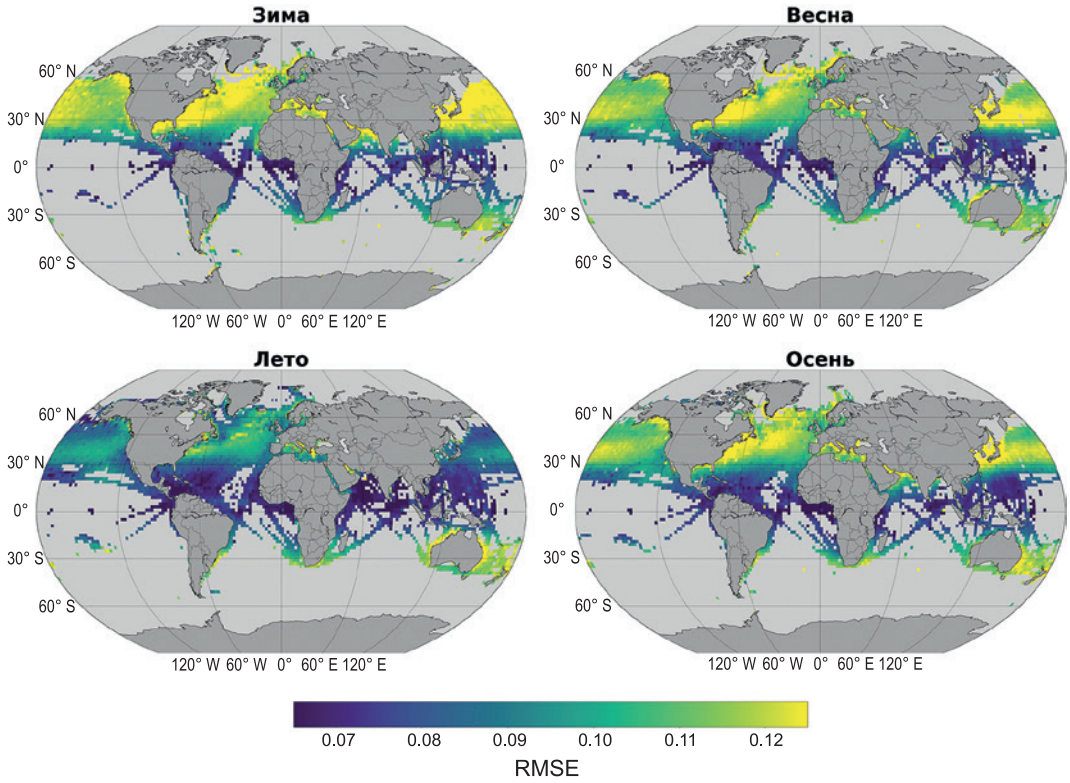

Рис. 9 – Карты ошибок тривиальной модели

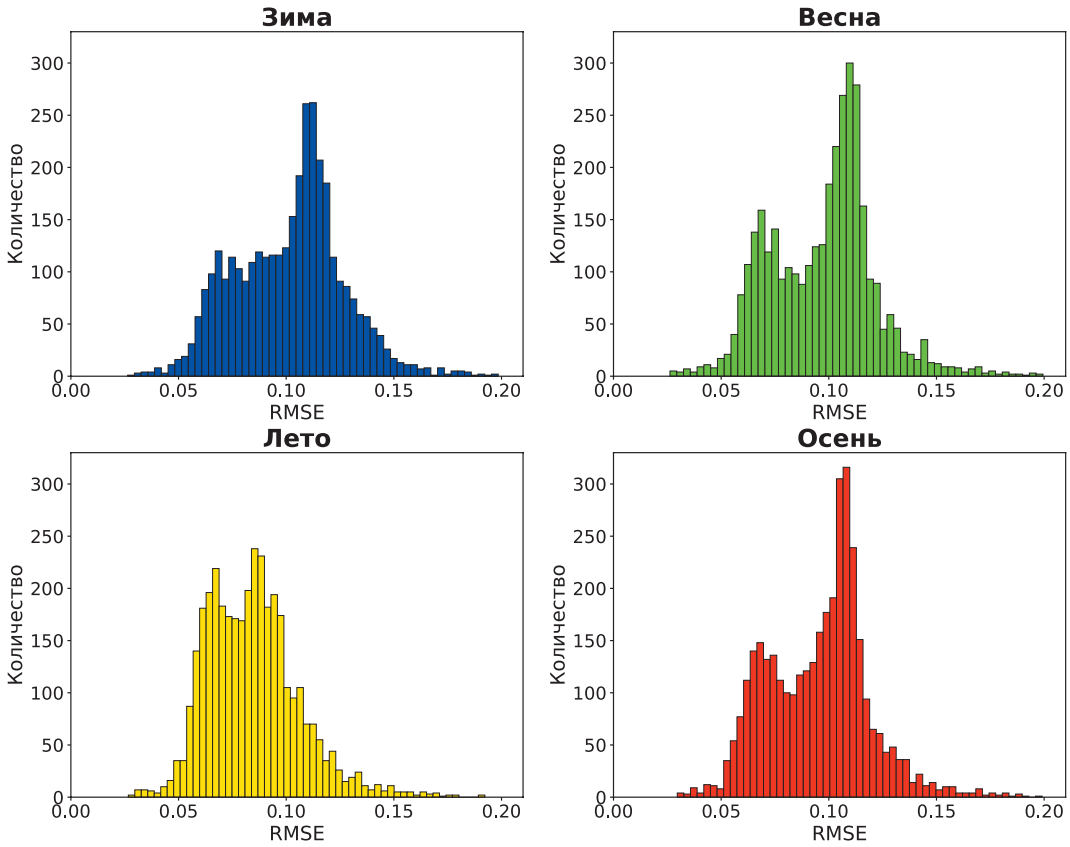

Рис. 10 – Гистограммы ошибок линейной регрессии

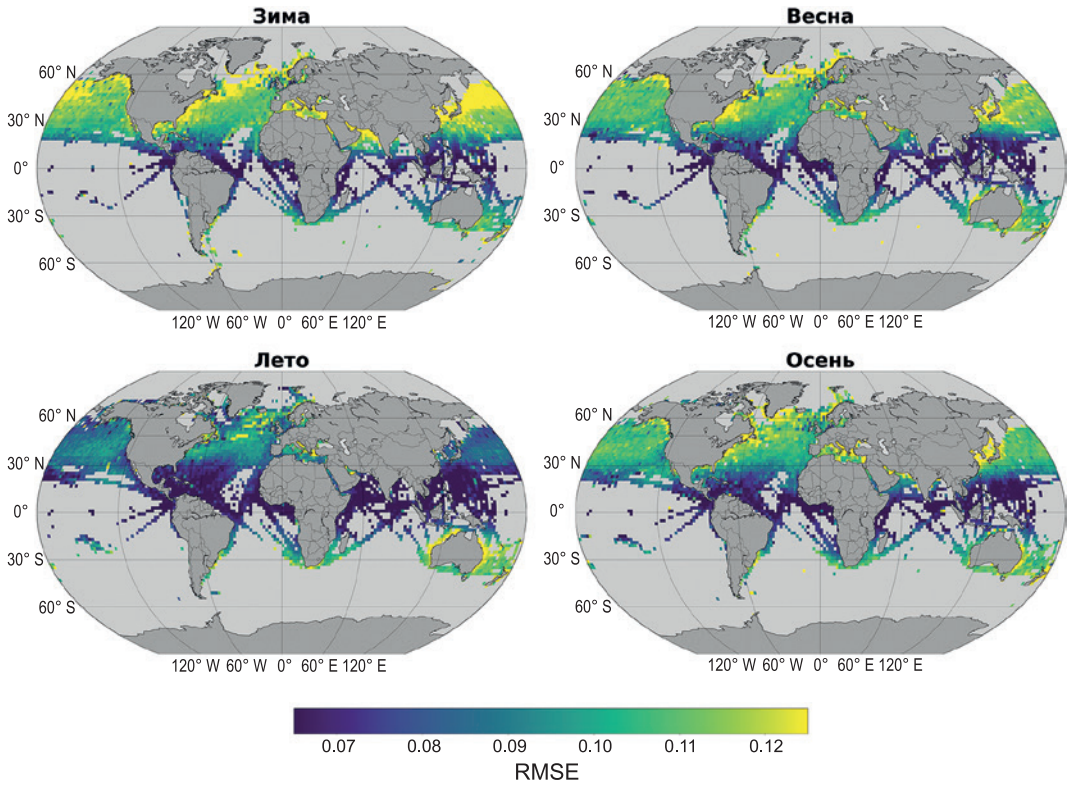

Рис. 11 – Карты ошибок линейной регрессии

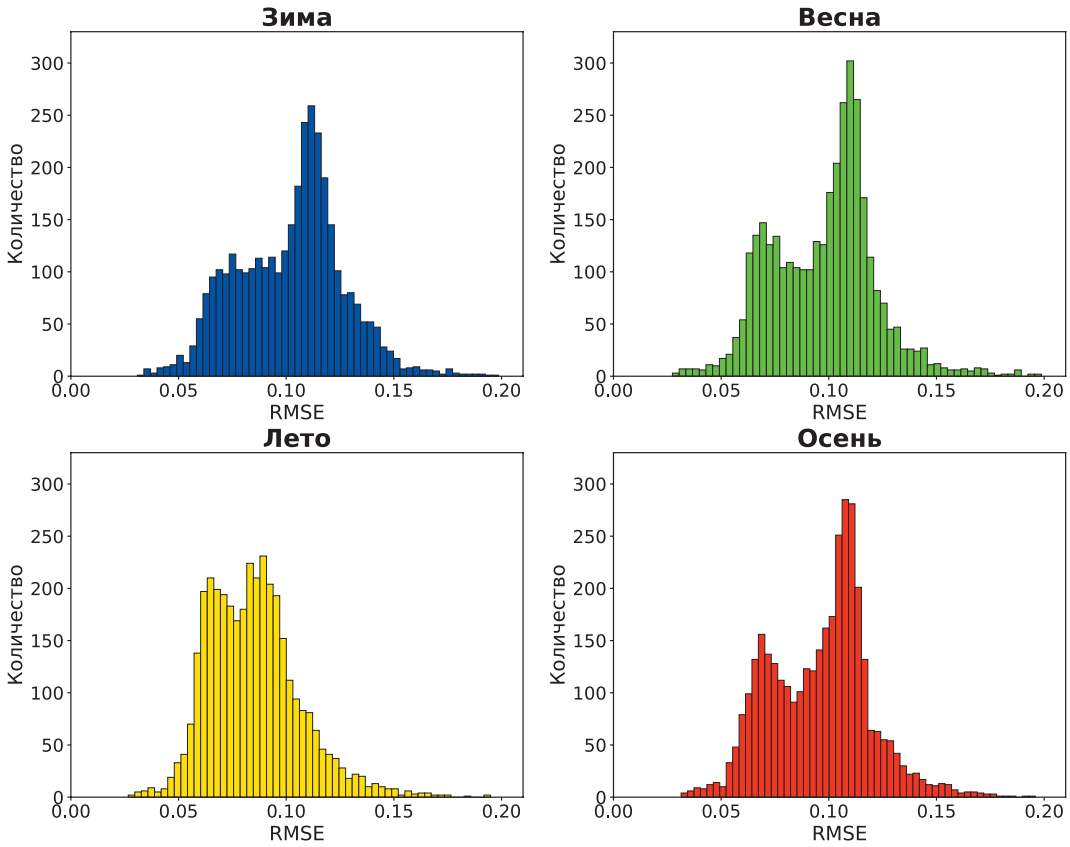

Рис. 12 – Гистограммы ошибок модели дерева решений

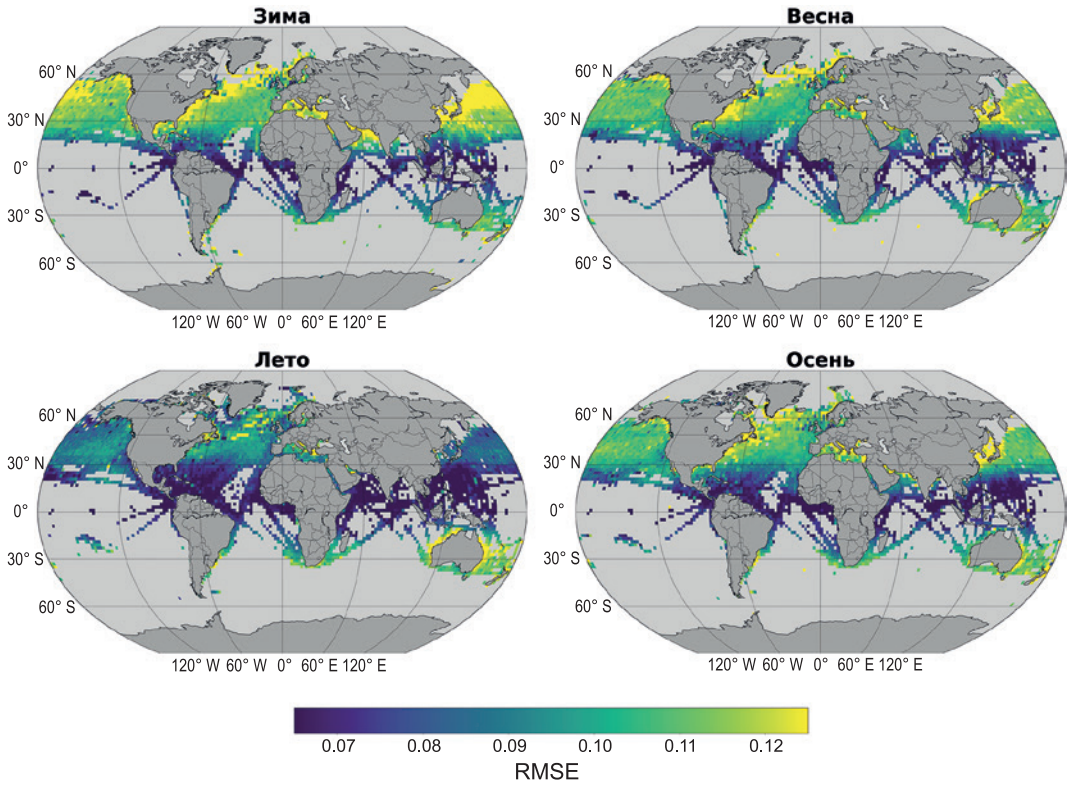

Рис. 13 – Карты ошибок модели дерева решений

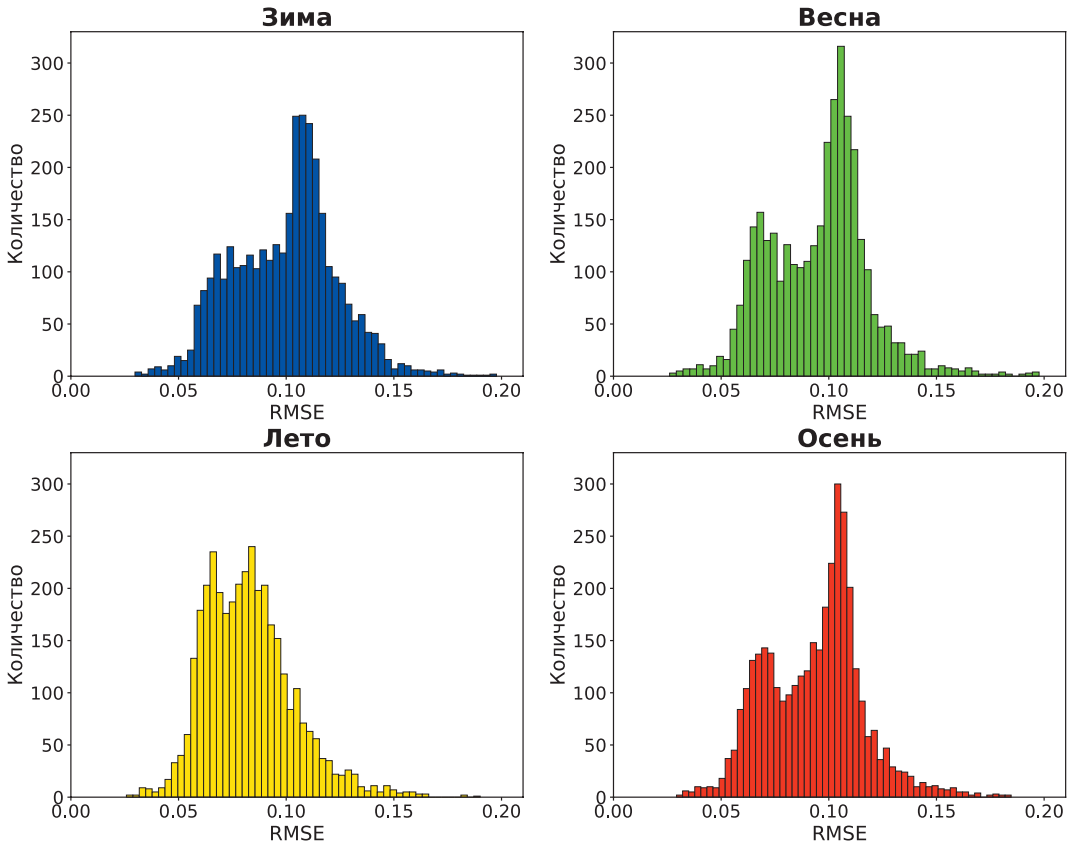

Рис. 14 – Гистограммы ошибок модели случайных лесов

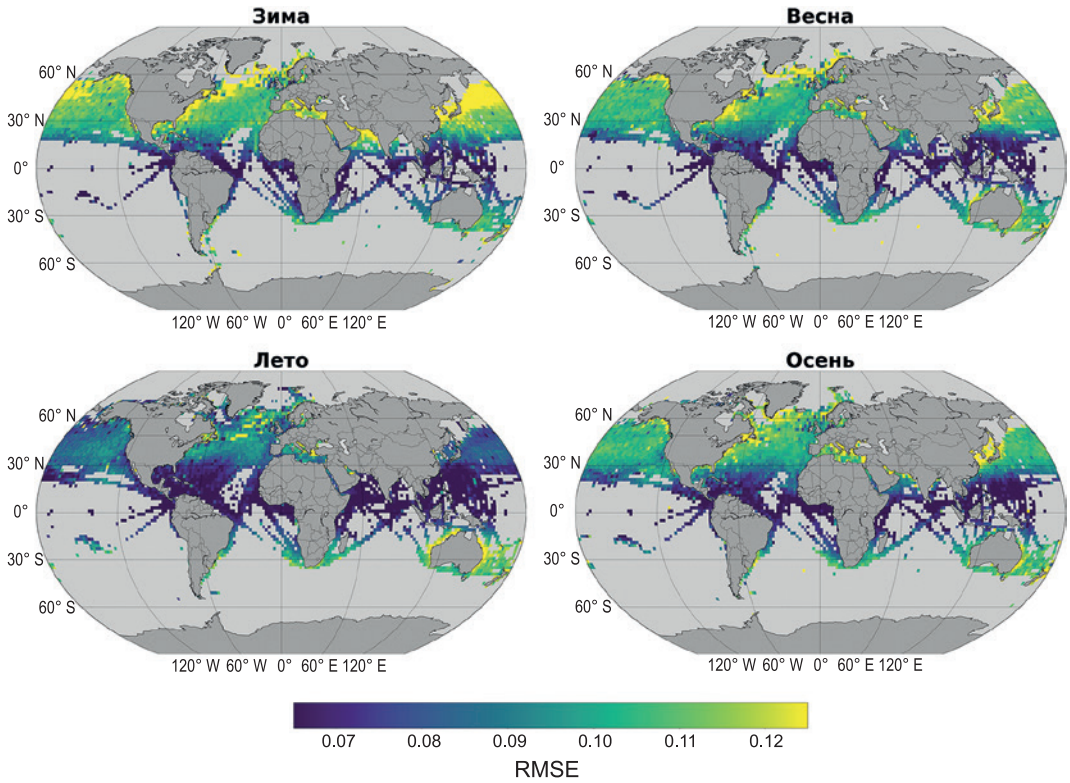

Рис. 15 – Карты ошибок модели случайных лесов

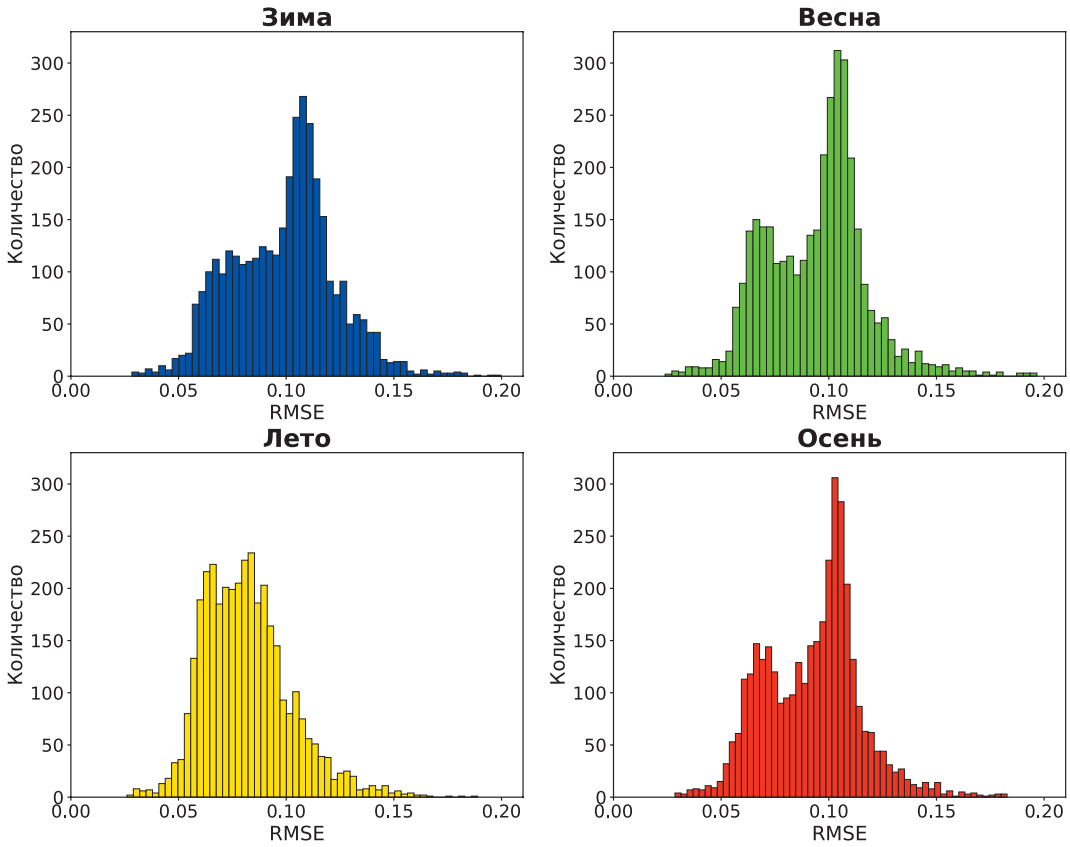

Рис. 16 – Гистограммы ошибок модели Catboost

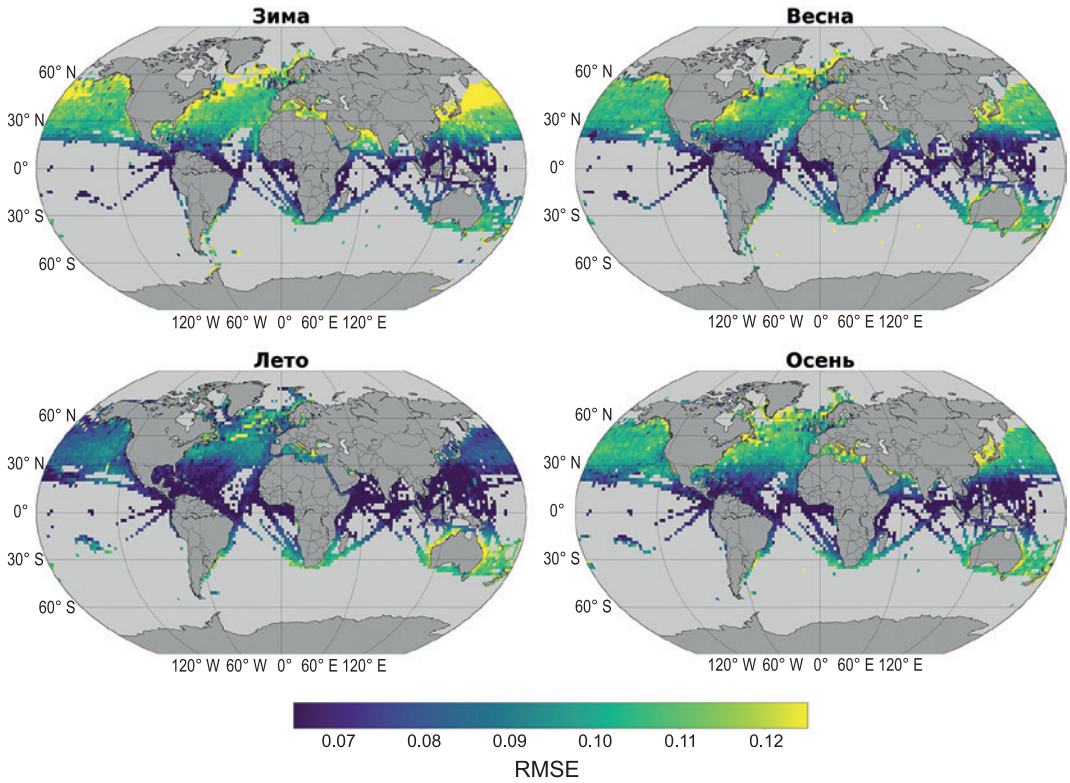

Рис. 17 – Карты ошибок модели Catboost

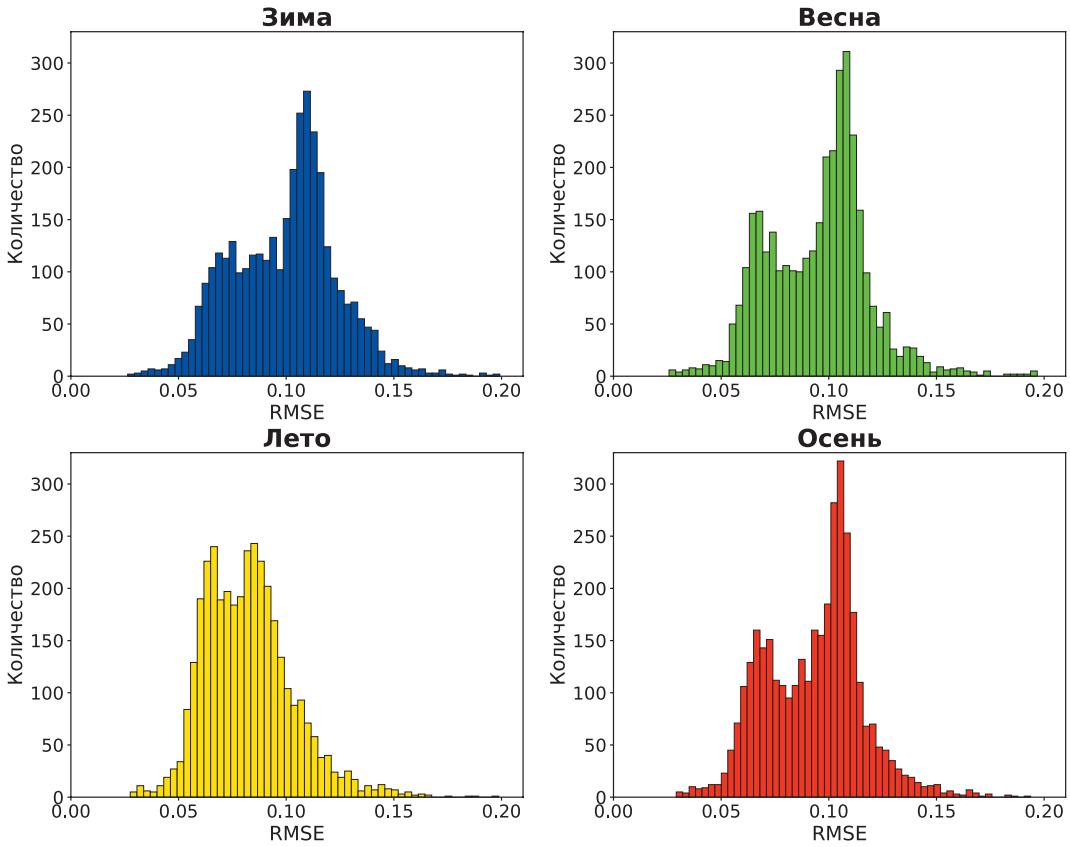

Рис. 18 – Гистограммы ошибок полносвязной искусственной нейронной сети

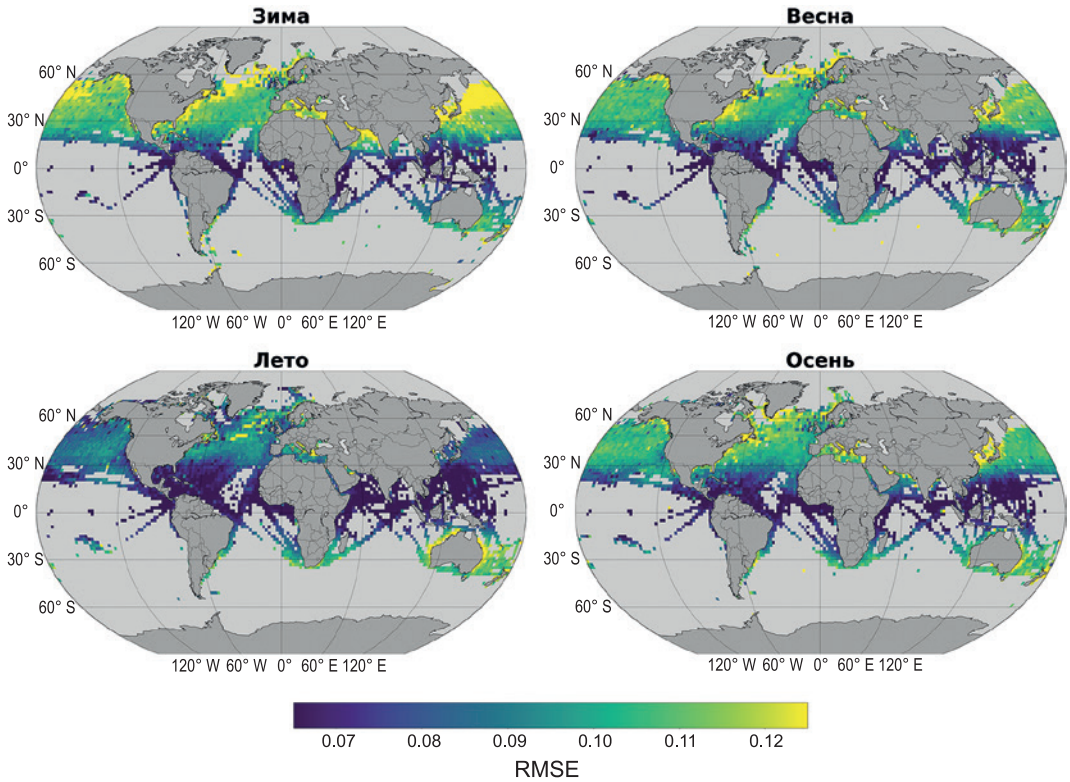

Рис. 19 – Карты ошибок полносвязной искусственной нейронной сети

Сравнение карт разброса относительной влажности и других карт климатических характеристик относительной влажности, построенных по данным DISO3 (унаследованным от коллекции ICOADS), с картами ошибок моделей позволяет сделать вывод, что качество моделей машинного обучения напрямую связано с изменчивостью относительной влажности. Большая вариабельность относительной влажности усложняет задачу, так как в регионах с ее выраженной изменчивостью наблюдается широкий диапазон значений, что затрудняет выявление устойчивых зависимостей между признаками (сопутствующими метеорологическими измерениями и наблюдениями) и целевой переменной. В таких условиях модели склонны ошибочно интерпретировать случайные отклонения как закономерности, что приводит к ухудшению качества восстановления относительной влажности.

Кроме того, высокая изменчивость влажности может указывать на сложные, нелинейные зависимости между признаками и целевой переменной. Линейные модели, такие как линейная регрессия, в силу заложенных в них предположений не подходят для их выявления. Применение более сложных моделей способно улучшить качество, но требует б***о***льших объемов данных для более точных результатов. В настоящем исследовании наблюдался недостаток данных в некоторых регионах, что существенно ухудшило точность аппроксимации относительной влажности.

Результаты показывают, что вблизи экватора, где изменчивость влажности минимальна, модели независимо от их сложности демонстрируют высокую точность. Однако с увеличением расстояния от экватора, где изменчивость возрастает, качество аппроксимации заметно ухудшается.

Сравнение карт ошибок моделей по сезонам также подтверждает эти выводы. В летний период, когда изменения влажности минимальны, даже простые модели показывают хорошие результаты. Напротив, зимой или в переходные сезоны, характеризующиеся высокой изменчивостью целевой переменной, ошибки аппроксимации значительно увеличиваются.

Эти результаты подчеркивают важность учета сезонных и географических особенностей при обучении моделей машинного обучения. В регионах с высокой изменчивостью влажности необходимы дополнительные меры для повышения точности, например, сбор большего объема данных или использование ансамблевых методов машинного обучения и нейросетей.

3.5. Сравнение моделей

Анализ ошибок показал, что полносвязная нейронная сеть и ансамблевые методы демонстрируют значительно лучшие результаты по качеству и стабильности по сравнению с тривиальной моделью и простыми алгоритмами.

Тривиальная модель, основанная на усреднении, игнорирует любые зависимости между признаками и целевой переменной, поэтому ее результаты подходят только для оценки базового уровня качества аппроксимации.

Линейная регрессия, хотя и учитывает линейные зависимости, ограничена в моделировании сложных, нелинейных взаимосвязей. В условиях высокой изменчивости влажности ее предположения о линейности приводят к большим ошибкам.

Дерево решений, в отличие от линейной регрессии, может учитывать нелинейные зависимости. Однако оно подвержено переобучению, особенно при ограниченном объеме данных, что делает его результаты менее стабильными.

Ансамблевые методы, такие как случайный лес и Catboost, обеспечивают более точные и стабильные результаты, поскольку они основаны на объединении множества деревьев. Случайный лес уменьшает переобучение, усредняя результаты множества деревьев, что повышает устойчивость модели к шуму и выбросам. Catboost улучшает результаты за счет последовательного уменьшения ошибок предыдущих деревьев.

Полносвязная нейронная сеть обеспечивает хорошее качество аппроксимации благодаря своей способности моделировать сложные, нелинейные зависимости. Архитектура искусственной нейронной сети позволяет эффективно учитывать как локальные, так и глобальные закономерности, что особенно важно при высокой изменчивости данных. Кроме того, нейронные сети способны обучаться на сложных, шумных данных, выявляя скрытые взаимосвязи.

4. Заключение

В этом исследовании были использованы модели машинного обучения для восстановления приповерхностной влажности атмосферы над океаном по данным сопутствующих метеорологических измерений.

Результаты работы подтверждают эффективность алгоритмов машинного обучения для восстановления данных относительной влажности по сопутствующим измерениям и наблюдениям и позволяют повысить точность восстановления климатических рядов по сравнению с тривиальными подходами, например, усреднением величины по регионам. Это подтверждается результатами экспериментов, в которых модели, такие как случайный лес, Catboost и полносвязная искусственная нейронная сеть, показали наилучшее качество, благодаря своей способности учитывать сложные нелинейные зависимости между признаками и целевой переменной.

Анализ результатов показал, что наибольшие ошибки аппроксимации наблюдаются в сезоны и в регионах с высокой изменчивостью влажности. Высокая вариабельность относительной влажности создает дополнительные сложности для аппроксимации. В таких условиях простые модели показывают низкое качество. В то же время в регионах, слабо обеспеченных данными наблюдений и измерений, даже более сложные модели, которые обладают существенной выразительной способностью и могут учитывать более широкий спектр зависимостей, сталкиваются с ограничениями. Примененный в работе подход, основанный на разделении данных по 2-градусным ячейкам и сезонам, позволил учесть региональные особенности и сезонные колебания влажности, а также снизил ошибки аппроксимации.

Анализ неопределенностей, проведенный с использованием подхода бутстреп, позволил оценить надежность полученных результатов и выявить наиболее значимые области для дальнейшего улучшения результатов. Одной из главных проблем, ограничивающих точность результатов, является недостаточное количество данных, что снижает качество аппроксимации и повышает неопределенность.

Сравнительный анализ моделей выявил, что Catboost и случайный лес продемонстрировали наилучшие результаты в данной работе, что делает их лучшими моделями для аппроксимации влажности. Catboost показал наивысшее среднее качество на отложенной выборке, а случайный лес показал наилучшее среднее качество по bootstrap-выборкам. Модели демонстрируют наиболее стабильные результаты с низкими значениями неопределенности как меры качества, так и целевой переменной, что подтверждает их способность к точному восстановлению относительной влажности.

Предложенный подход восстановления относительной влажности по данным сопутствующих измерений и наблюдений может быть в дальнейшем использован для восполнения недостающих значений в климатических рядах относительной влажности атмосферы над океаном. Это позволит расширить период климатических исследований, связанных с относительной влажностью, в прошлое на период наличия данных, представленных в ICOADS.

Перспективы дальнейших исследований включают несколько приоритетных направлений, способных существенно улучшить точность восстановления относительной влажности. Одним из важных направлений является разработка и внедрение новых моделей глубокого обучения, способных лучше учитывать сложные пространственно-временные зависимости. Не менее значимым направлением является использование дополнительных источников данных, например, реанализов. Еще одним перспективным направлением является разработка промежуточных моделей, которые оценивают дополнительные параметры и преобразуют их в новые признаки для обучения. Кроме того, попытки использовать темпоральные данные в рамках серии наблюдений в экспедициях также могут повысить точность, так как они позволят моделям улавливать изменения во времени, что особенно важно для анализа климатических процессов.

Благодарности. Исследование выполнено в рамках Соглашения № 075-03-2025-662 от 17.01.2025 Московского физико-технического института. Оптимизация гиперпараметров моделей машинного обучения проведена при поддержке государственного задания FMWE-2024-0017 Института океанологии им. П. П. Ширшова РАН.

Список литературы

1. *Александрова М. П., Гулев С. К.* Реконструкция долгопериодной изменчивости облачности и радиационных потоков над Мировым океаном: Дис. на канд. физико-математических наук: 01.06.17. М.: ИО РАН, 2023.
2. *Adele Cutler, D. Richard Cutler, John R. Stevens* Random Forests // J. Springer Nature. Ensemble Machine Learning. 2012. https://doi.org/10.1007/978-1-4419-9326-7_5.
3. *Barry de Ville* Decision trees // J. WIREs Computational Statistics. 2013. Vol. 5. Iss. 6. https://doi.org/10.1002/wics.1278.
4. *Callum J. Shakespeare, Michael L. Roderick* What Controls Near-Surface Relative Humidity Over the Ocean // James. Journal of Advances in Modeling Earth System. 2024. https://doi.org/10.1029/2023MS004168.
5. *Carsten Abraham, Colin Goldblatt* Changes in Relative Humidity Profiles over Earth's Oceans in a Warming Climate: A Satellite-Data-Based Inference // Journal of the Atmospheric Sciences. 2023. https://doi.org/10.1175/JAS-D-22-0119.1.
6. *Efron B.* Bootstrap Methods: Another Look at the Jackknife // The Annals of Statistics. 1979. No. 1 (7). P. 1–26.
7. *Eric Freeman, Scott D. Woodruff, Steven J. Worley, Sandra J. Lubker, Elizabeth C. Kent, William E. Angel, David I. Berry, Philip Brohan, Ryan Eastman, Lydia Gates, Wolfgang Gloeden, Zaihua Ji, Jay Lawrimore, Nick A. Rayner, Gudrun Rosenhagen, Shawn R. Smith* ICOADS Release 3.0: a major update to the historical marine climate record // International Journal of Climatology. 2017. No. 5 (37). https://doi.org/10.1002/joc.4775.
8. *Hancock J. T., Khoshgoftaar T. M.* CatBoost for big data: an interdisciplinary review // J. Big Data 7. 2020. https://doi.org/10.1186/s40537-020-00369-8.
9. *Hussain Almarzooq, Umair bin Waheed* Automating hyperparameter optimization in geophysics with Optuna: A comparative study // J. Geophysical Prospecting. 2024. https://doi.org/10.1111/1365-2478.13484.
10. *Jiajun Guo, Liang Zhang, Ruqiang Guo* Relative humidity prediction with covariates and error correction based on SARIMA-EG-ECM model // J. Springer Nature. Modeling Earth Systems and Environment. 2023. https://doi.org/10.1007/s40808-023-01738-x.
11. *John R. Borchert* Regional Differences in the World Atmospheric Circulation // J. Annals of the Association of American Geographers. 2008. https://doi.org/10.2307/2561080.
12. *Luminda Niroshana Gunawardhana, Ghazi A. Al-Rawas, So Kazama* An alternative method for predicting relative humidity for climate change studies // J. Meteorological application. 2017. https://doi.org/10.1002/met.1641.
13. *Marvin Minsky, Seymour Papert* Perceptrons: An Introduction to Computational Geometry // The MIT Press. 1969.
14. *Matthieu Parizy, Norihiro Kakuko, Nozomu Togawa* Fast Hyperparameter Tuning for Ising Machines // 2023 IEEE International Conference on Consumer Electronics (ICCE). 2023. https://doi.org/10.1109/ICCE56470.2023.10043382.
15. *Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Kayama* Optuna: A Next-generation Hyperparameter Optimization Framework // Conference: the 25th ACM SIGKDD International Conference. 2019. https://doi.org/10.48550/arXiv.1907.10902.

Статья поступила в редакцию 03.02.2025, одобрена к печати 12.05.2025.
Для цитирования: *Вострикова С. А., Криницкий М. А., Гулев С. К., Александрова М. П.* Восстановление приповерхностной влажности атмосферы над океаном по данным сопутствующих метеорологических измерений с применением методов машинного обучения // Океанологические исследования. 2025. № 53 (2). С. 121–148. https://doi.ocean.ru/10.29006/1564-2291.JOR-2025.53(2).7.

ISSN (online): 2587-9634 / ISSN (print): 1564-2291
Океанологические исследования. 2025. Том 53. № 2. С. 121–148

RECONSTRUCTION OF ATMOSPHERIC SURFACE HUMIDITY OVER THE OCEAN FROM RELEVANT METEOROLOGICAL MEASUREMENTS USING MACHINE LEARNING METHODS

S. A. Vostrikova[1], M. A. Krinitsky[1,2], S. K. Gulev[2], M. P. Alexandrova[2]

[1] Moscow Institute of Physics and Technology,
9, Institutskiy pereulok, Dolgoprudny, 117303, Russia,
e-mail: vostrikova.sa@phystech.edu;
[2] Shirshov Institute of Oceanology, Russian Academy of Sciences,
36, Nakhimovskiy prospekt, Moscow, 117997, Russia,

Air humidity in the near-surface layer of the atmosphere over the ocean is a key climate parameter that has a significant impact on the processes of moisture and heat transfer between the ocean and the atmosphere, as well as on the dynamics of atmospheric processes in general. Analysis of meteorological data collected during the 20th century shows the sparseness of humidity measurement series in space and time. The International Ocean and Atmosphere Data Set (ICOADS) indicates an insufficient density of measurements in the early 20th century compared to later periods, which creates difficulties for adequate analysis of climate trends in relative humidity.

Methods for approximating humidity time series presented in the literature often demonstrate limited accuracy, based mainly on statistical and heuristic approaches. Our work is aimed at improving the quality of solving this problem through the use of machine learning methods.

As a first, simplest approach, we solved the problem in the formulation of the approximation of relative humidity based on the data of accompanying measurements of atmospheric pressure, air temperature, wind speed and direction, ocean surface temperature, as well as observations of the amount and types of clouds at three tiers. In addition, the accompanying variables include the WMO weather code and the estimated solar altitude. The study used the following types of machine learning models: linear regression, decision tree, random forest, gradient boosting, and fully connected artificial neural network. To improve the territorial and temporal specificity of the developed models, we conducted a study for each 2-degree square and each season separately. The scikit-learn library and the package implementing the CatBoost model were used to train and apply the machine learning models. For each type of model, we optimized the hyperparameters using the Optuna Bayesian optimization library. Based on the results obtained, maps of the spatial distribution of model errors were constructed, which made it possible to identify regions with high and low accuracy of humidity approximation.

The study confirmed the effectiveness of machine learning methods for reconstructing climate series, identified the most suitable models for this task, and outlined promising areas for further work.

Keywords: relative humidity, data reconstruction, machine learning, regression, linear regression, decision tree, random forest, gradient boosting, multilayer perceptron

Acknowledgement: The study was supported by the Agreement No. 075-03-2025-662 dated 17.01.2025 of the Moscow Institute of Physics and Technology. Hyperparameters optimization was supported by the government contract FMWE-2024-0017 of the Shirshov Institute of Oceanology of the Russian Academy of Sciences.

References

1. Adele Cutler, D. Richard Cutler, and R. John, 2012: Stevens Random Forests *J. Springer. Ensemble Machine Learning.* https://doi.org/10.1007/978-1-4419-9326-7_5.
2. Aleksandrova, M. P. and S. K. Gulyov, 2017: *Rekonstrukciya dolgoperiodnoj izmenchivosti oblachnosti i radiacionnyh potokov nad mirovym okeanom (Reconstruction of long-term variability of cloudiness and radiation fluxes over the world ocean).* Dis. na kand. fiziko-matematicheskih nauk. 01.06.2017. Moscow, IO RAN.
3. Barry de Ville, 2013: Decision trees *J. WIREs Computational Statistics.* **5** (6), https://doi.org/10.1002/wics.1278.
4. Callum J. Shakespeare and Michael L. Roderick, 2024: What Controls Near-Surface Relative Humidity Over the Ocean *J. James. Journal of Advances in Modeling Earth System.* https://doi.org/10.1029/2023MS004168.
5. Carsten Abraham and Colin Goldblatt, 2023: Changes in Relative Humidity Profiles over Earth's Oceans in a Warming Climate: A Satellite-Data-Based Inference. *Journal of the Atmospheric Sciences.* https://doi.org/10.1175/JAS-D-22-0119.1.
6. Efron, B., 1979: Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **1** (7), 1–26.
7. Freeman, Eric, Scott D. Woodruff, Steven J. Worley, Sandra J. Lubker, Elizabeth C. Kent, William E. Angel, David I. Berry, Philip Brohan, Ryan Eastman, Lydia Gates, Wolfgang Gloeden, Zaihua Ji, Jay Lawrimore, Nick A. Rayner, Gudrun Rosenhagen, and Shawn R. Smith, 2017: ICOADS Release 3.0: a major update to the historical marine climate record. *International Journal of Climatology,* **5** (37), https://doi.org/10.1002/joc.4775.
8. Hancock, J. T. and T. M. Khoshgoftaar, 2020: CatBoost for big data: an interdisciplinary review. *J. Big Data 7,* https://doi.org/10.1186/s40537-020-00369-8.
9. Hussain Almarzooq, Umair bin Waheed, 2024: Automating hyperparameter optimization in geophysics with Optuna: A comparative study. *J. Geophysical Prospecting,* https://doi.org/10.1111/1365-2478.13484.
10. Jiajun Guo, Liang Zhang, and Ruqiang Guo, 2023: Relative humidity prediction with covariates and error correction based on SARIMA-EG-ECM model *J. Springer Nature. Modeling Earth Systems and Environment,* https://doi.org/10.1007/s40808-023-01738-x.
11. John, R., 2008: Borchert Regional Differences in the World Atmospheric Circulation *J. Annals of the Association of American Geographers,* https://doi.org/10.2307/2561080.
12. Luminda Niroshana Gunawardhana, Ghazi A. Al-Rawas, So Kazama, 2017: An alternative method for predicting relative humidity for climate change studies. *J. Meteorological application,* https://doi.org/10.1002/met.1641.
13. Marvin Minsky, Seymour Papert, 1969: Perceptrons: An Introduction to Computational Geometry *The MIT Press.*
14. Matthieu Parizy, Norihiro Kakuko, Nozomu Togawa Fast Hyperparameter, 2023: Tuning for Ising Machines *2023 IEEE International Conference on Consumer Electronics (ICCE),* https://doi.org/10.1109/ICCE56470.2023.10043382.
15. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Kayama Optuna, 2019.: A Next-generation Hyperparameter Optimization Framework. *Conference: the 25th ACM SIGKDD International Conference.* https://doi.org/10.48550/arXiv.1907.10902.

Submitted 03.02.2025, accepted 12.05.2025.
For citation: Vostrikova, S. A., M. A. Krinitsky, S. K. Gulev, and M. P. Alexandrova, 2025: Reconstruction of atmospheric surface humidity over the ocean from relevant meteorological measurements using machine learning methods. *Journal of Oceanological Research*, **53** (2), 121–148, https://doi.ocean.ru/10.29006/1564-2291.JOR-2025.53(2).7.