

Восстановление пропусков в многомерных временных рядах методами математической статистики и машинного обучения

В. С. Алешновский,^{1,*} В. К. Авилов,² П. В. Голубцов,¹ Ю. А. Курбатова,²
 О. А. Куричева,² Д. А. Макеев,¹ П. А. Фадеева,¹ А. И. Чуличков,^{1,3} Н. Е. Шапкина^{1,†}
¹Московский государственный университет имени М. В. Ломоносова, физический факультет
 Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2
²Институт проблем экологии и эволюции имени А. Н. Северцова РАН
 Россия, 119071, Москва, Ленинский пр-т., д. 33
³Институт физики атмосферы имени А. М. Обухова РАН
 Россия, 119017, Москва, Пыжевский пер., д. 3

(Поступила в редакцию 09.04.2026; после доработки 03.05.2026; подписана в печать 04.05.2026)

Объектом исследования являются многомерные временные ряды геофизических параметров. В ходе работы применялись методы статистики и машинного обучения для восстановления пропусков в данных. Были достигнуты следующие результаты: для ряда концентрации CO₂ на высоте 30 м модель ARIMAX показала наименьшие значения метрик RMSE и MAE, затратив всего 2.3 с на восстановление, что значительно быстрее по сравнению с моделями нейронных сетей LSTM (251 с) и GRU (160 с). Восстановление данных температуры воздуха наиболее эффективно выполнено с использованием рекуррентной нейронной сети GRU. Результаты показывают, что модели семейства ARIMA являются более подходящими для восстановления данных при наличии коррелированных параметров, тогда как в других случаях следует отдавать предпочтение моделям на основе нейронных сетей GRU. Выводы могут быть использованы для улучшения методик восстановления данных в многомерных временных рядах.

PACS: 02.70.Rr, 02.50.Ey УДК: 519.21, 519.25

Ключевые слова: многомерные временные ряды, геофизические параметры, восстановление пропусков данных, математическое моделирование, диоксид углерода.

[10.55959/MSU0579-9392.81.2630101](https://doi.org/10.55959/MSU0579-9392.81.2630101)

ВВЕДЕНИЕ

Из-за возрастающего влияния метеорологических факторов на окружающую среду, особенно в связи с увеличением концентрации углекислого газа (CO₂), который играет важную роль в фотосинтезе и формировании климата, актуальность исследований в этой области возрастает. Рост концентрации CO₂ вызван активным использованием ископаемых энергоносителей и обезлесением, и его изучение критично для понимания и прогнозирования изменений в атмосфере и климате [1, 2]. Ключевыми задачами климатических и экологических исследований являются получение высококачественных данных и анализ современных изменений метеорологических параметров атмосферы. Современные технологии позволяют научному сообществу использовать автоматические комплексы для мониторинга широкого спектра абиотических факторов, включая концентрации и потоки парниковых газов и метеорологические показатели [3]. Однако в процессе измерений могут возникать пропуски данных из-за технических проблем, что делает важным задачу восстановления пропус-

ков в многомерных временных рядах с помощью математического моделирования для обеспечения непрерывности регистрации параметров [4, 5]. Многомерные временные ряды — это наборы данных, в которых каждое наблюдение содержит информацию о нескольких переменных, собранных в разные моменты времени [6]. В результате, подобные исследования способствуют оценке динамики процессов, определяющих климаторегулирующие функции экосистем.

Цель данной работы заключается в создании и сравнении математических моделей, основанных на статистических подходах и методах машинного обучения, с целью восстановления рядов динамики при помощи нескольких входных параметров. Исследование проводилось на основе данных, собранных в ходе наблюдений за концентрацией углекислого газа и температурой воздуха на эколого-климатической станции AsiaFlux, расположенной в тропическом муссонном лесу южного Вьетнама, в период с 2011 по 2020 гг. Полученные результаты имеют важное значение для исследования реакции тропических лесов на текущие изменения в глобальном климате и в первую очередь применимы к экосистемам с аналогичными климатическими условиями. Для других климатических зон, отличающихся сезонной структурой и статистическими характеристиками временных рядов, требуется дополнительная валидация предложенных подходов.

* E-mail: aleshnovskii.vs17@physics.msu.ru

† E-mail: neshapkina@mail.ru

1. ПРОБЛЕМАТИКА ИССЛЕДОВАНИЯ

Восстановление пропусков в многомерных геофизических временных рядах остаётся сложной задачей, несмотря на развитый аппарат анализа временных рядов. Основная трудность заключается в гетерогенности динамики разных переменных: одни ряды имеют чёткий тренд и выраженную сезонность, другие демонстрируют резкие нелинейные изменения. Универсальный метод, одинаково хорошо описывающий столь разнородные процессы, практически отсутствует, а применение единого подхода без учёта физической специфики каждой переменной приводит к потере точности и физической адекватности построенной модели.

Также важно учесть многомерность данных и взаимозависимость переменных. В климатической системе температура, концентрация парниковых газов, турбулентность и радиационный баланс связаны нелинейными кросс-корреляциями и причинно-следственными связями с запаздыванием. Игнорирование этих взаимосвязей при использовании традиционных одномерных методов означает упущение значительной части доступной информации и снижение качества восстановления [7, 8]. При этом можно столкнуться с классической дилеммой: параметрические статистические модели (ARIMA, ARIMAX и др.) интерпретируемы, экономичны и устойчивы, но плохо справляются со сложными нелинейностями; современные нейросетевые подходы (LSTM, GRU, трансформеры) гораздо более гибкие, но требуют больших объёмов данных, склонны к переобучению, вычислительно затратны и часто непрозрачны («чёрный ящик»).

В прикладных задачах экологического мониторинга и потоковой обработки данных дополнительно возникает потребность в хорошем балансе точности и вычислительной эффективности. Существующие методы не всегда могут дать решения, которое одновременно учитывало бы физическую природу переменных, структуру межпеременных связей, компромисс между интерпретируемостью и моделирующей мощностью, а также ограничения по времени расчёта. Это и определяет актуальность поиска новых или гибридных подходов к восстановлению в многомерных геофизических временных рядах.

2. ПОСТАНОВКА ЗАДАЧИ

Задача восстановления пропусков в многомерных временных рядах в определенном смысле эквивалентна задаче прогнозирования: в обоих случаях требуется восстановить неизвестные значения на основе имеющихся наблюдений. Это позволяет применять для заполнения пропусков мощные модели, разработанные для прогнозирования (ARIMAX, LSTM, GRU). Исследование проводится на реальных многомерных данных со станции AsiaFlux с искусственно созданными пропусками.

Рассматриваются следующие задачи.

1. Адаптировать и реализовать два класса методов для восстановления пропусков: статистические — модели ARIMAX с учётом экзогенных переменных, машинное обучение — рекуррентные нейронные сети LSTM и GRU.
2. Выполнить сравнительное тестирование моделей по двум группам критериев: точность — RMSE, MAE, r на отложенной тестовой выборке; эффективность — время обучения модели.
3. Проанализировать результаты с математической и физической точек зрения: определить, для каких переменных предпочтительны статистические, а для каких — нейросетевые подходы.

На основе анализа сформулировать практические рекомендации по выбору метода заполнения пропусков.

3. МЕТОДЫ ВОССТАНОВЛЕНИЯ РЯДОВ

3.1. ARIMA

ARIMA (авторегрессионная интегрированная модель скользящего среднего) является статистической моделью, применяемой для анализа и прогнозирования временных рядов [9, 10]. Эта модель включает в себя три составляющие: авторегрессию (AR), интеграцию (I) и скользящее среднее (MA), как показано в формуле 1 [11]:

$$\Delta^d y_t = \sum_{i=1}^p \phi_i \Delta^d y_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad (1)$$

где p — порядок авторегрессии, d — порядок интегрируемости (разности), q — порядок скользящего среднего, ϵ_t — случайная ошибка (остаток модели), предполагаемая нормально распределённой, ϕ_i , θ_j — параметры модели, а Δ^d — оператор разности временного ряда порядка d .

- Авторегрессия (AR) описывает зависимость текущих значений временного ряда от предшествующих значений. Параметр p в модели ARIMA определяет количество предыдущих значений, используемых для прогнозирования последующих значений.
- Интеграция (I) служит для преобразования нестационарных рядов в стационарные с целью прогнозирования. Параметр d в модели ARIMA обозначает количество интеграций, необходимых для превращения ряда в стационарный.

- Скользящее среднее (МА) применяется для устранения шума и сглаживания ряда. Параметр q в модели *ARIMA* определяет количество значений скользящего среднего, применяемых для прогнозирования будущих значений.

Таким образом, параметры *ARIMA* (p, d, q) характеризуют порядок авторегрессии, порядок интеграции и порядок скользящего среднего соответственно.

3.2. ARIMAX

Модель *ARIMAX* представляет собой расширение классической модели *ARIMA*, которое позволяет учитывать влияние внешних факторов на целевой временной ряд (формула 2) [12]:

$$\Delta^d y_t = \sum_{i=1}^p \phi_i \Delta^d y_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \sum_{m=1}^M \beta_m X_{m,t} + \epsilon_t, \\ \epsilon_t \sim N(0, \sigma^2), \quad (2)$$

где β_m — коэффициенты регрессии при M внешних (экзогенных) переменных.

Экзогенные переменные ($X_{m,t}$) — это внешние факторы, значения которых задаются вне модели и не объясняются внутри неё самой. Они считаются внешними по отношению к моделируемому процессу и обычно рассматриваются как заданные (известные) на момент прогнозирования. В контексте метеорологических данных это, например, другие измеряемые показатели на станции (температура, влажность, давление и т.д.), если целевой ряд — это, скажем, концентрация углекислого газа.

В отличие от них *эндогенные переменные* — это переменные, которые *определяются внутри самой модели* и объясняются её структурой (в случае *ARIMAX* это сам ряд y_t и его лаги/прошлые ошибки). Иными словами,

- эндогенная переменная $\leftarrow$ зависит от самой себя в прошлом и от ошибок модели;
- экзогенная переменная $\leftarrow$ приходит извне и влияет на эндогенную, но сама по себе в модели не объясняется.

В рассматриваемом случае другие показатели, измеряемые на метеорологической станции, выступают именно в роли экзогенных переменных [13].

Экзогенные переменные (X) могут представлять собой самые разные факторы, которые предположительно оказывают влияние на целевой временной ряд и включаются в модель в качестве дополнительных предикторов [14]. Их правильный выбор и включение часто существенно повышает качество прогноза по сравнению с обычной *ARIMA*.

3.3. Долгая краткосрочная память (LSTM)

Сеть долгой краткосрочной памяти (LSTM) — это разновидность рекуррентной нейронной сети, которая хорошо подходит для прогнозирования временных рядов с долгосрочными зависимостями [15, 16].

Модель обрабатывает данные шаг за шагом. На каждом временном шаге она получает текущие значения всех переменных (вектор $\mathbf{x}_t$) и информацию из предыдущего шага (скрытое состояние h_{t-1}).

LSTM использует специальные управляющие механизмы (вентили забывания, ввода и вывода), которые решают:

- какую часть прошлой информации сохранить,
- какую новую информацию добавить,
- какую часть результата передать дальше.

В итоге на каждом шаге формируется обновлённое скрытое состояние h_t , которое накапливает важные сведения из всей предыдущей последовательности.

Для получения прогноза на горизонт N шагов вперёд используется линейный выходной слой, который выдаёт сразу вектор предсказаний:

$$\hat{\mathbf{y}}_{t+1:t+N} = W_y h_t + b_y,$$

где $\hat{\mathbf{y}}_{t+1:t+N}$ — вектор прогнозов размером N , W_y — матрица весов размером $[N \times \text{размер скрытого состояния}]$, b_y — вектор смещений размером N .

В многомерных временных рядах модель автоматически учитывает взаимосвязи между переменными, что позволяет эффективно восстанавливать пропуски и прогнозировать на средне- и долгосрочный горизонт в сложных геофизических данных [17].

3.4. Управляемый рекуррентный блок (GRU)

Управляемый рекуррентный блок (GRU) — это упрощённая и более быстрая версия LSTM, которая тоже хорошо справляется с долгосрочными зависимостями в временных рядах [18, 19].

GRU работает похожим образом: на каждом шаге получает текущие значения переменных x_t и предыдущее состояние h_{t-1} .

Модель использует два управляющих механизма (вентиль обновления и вентиль сброса), которые определяют:

- насколько сильно заменить старую информацию новой,
- какую часть прошлой информации использовать для расчёта нового кандидата.

В результате получается новое скрытое состояние h_t , которое сочетает старую и новую информацию.

Прогноз на горизонт N шагов вперёд формируется тем же линейным слоем, что и в случае LSTM (см. формулу выше).

По сравнению с LSTM, GRU требует меньше вычислений, но часто показывает близкую точность, что делает её удобной для задач восстановления пропусков и прогнозирования в больших объёмах данных.

4. МЕТРИКИ ОЦЕНКИ КАЧЕСТВА МОДЕЛЕЙ

Для сравнения качества построенных моделей использовались стандартные метрики точности прогнозирования, рассчитанные на отложенной тестовой выборке. Тестовая выборка представляла собой часть реального временного ряда, которая полностью исключалась из процесса обучения и подбора гиперпараметров. Таким образом, на тестовом периоде модели делали прогноз на основе обученных параметров и информации из обучающей части (ARIMA); при этом для моделей с экзогенными переменными (ARIMAX, LSTM, GRU) дополнительно использовались реальные значения экзогенных переменных из тестовой выборки, известные на момент прогноза. При этом итоговый прогноз $\hat{y}_t$ вычислялся как среднее по нескольким независимым запускам обучения, что позволяет снизить влияние случайного стартового состояния на оцениваемые метрики. Полученные предсказания сравнивались с фактическими значениями целевой переменной.

4.1. Метрики точности прогноза

- **RMSE** (Root Mean Squared Error) — корень из среднеквадратичной ошибки:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}, \quad (3)$$

где y_t — фактическое значение, $\hat{y}_t$ — предсказанное значение, n — длина тестовой выборки.

Показатель RMSE выражается в тех же единицах измерения, что и целевая переменная, и за счёт квадратирования отклонений придаёт больший вес крупным ошибкам по сравнению с малыми.

- **MAE** (Mean Absolute Error) — средняя абсолютная ошибка:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|. \quad (4)$$

MAE показывает среднее по модулю отклонение прогноза от реальных данных и более устойчива к выбросам, чем RMSE.

- **r** (коэффициент корреляции Пирсона) — мера линейной связи между фактическими и прогнозными значениями:

$$r = \frac{\sum_{t=1}^n (y_t - \bar{y})(\hat{y}_t - \bar{\hat{y}})}{\sqrt{\sum_{t=1}^n (y_t - \bar{y})^2} \cdot \sqrt{\sum_{t=1}^n (\hat{y}_t - \bar{\hat{y}})^2}}, \quad (5)$$

где $\bar{y}$ — среднее фактических значений, $\bar{\hat{y}}$ — среднее прогнозных значений на тестовой выборке.

Здесь r принимает значения от -1 до 1 : значение, близкое к 1 , свидетельствует о сильной положительной линейной связи между прогнозом и фактом; значение около 0 — об отсутствии линейной зависимости; отрицательные значения указывают на обратную связь, что при прогнозировании означает систематически неверное направление предсказания.

Указанные метрики взаимодополняющие: RMSE и MAE характеризуют абсолютную величину ошибки, а r показывает, насколько прогноз связан с фактическими значениями.

4.2. Эффективность (вычислительная сложность)

Время обучения и время восстановления модели (t_{learn}) оценивались отдельно, потому что их приемлемость сильно зависит от временного масштаба самих данных и сценария применения.

- Для высокочастотных данных (секундные, минутные измерения) модель, которая обучается часы или даже десятки минут на новых поступающих данных, становится практически неприменимой — она просто не успевает обновляться с той же скоростью, с которой приходят новые наблюдения.
- Для низкочастотных данных (суточные, недельные измерения) даже обучение в несколько минут или десятков минут остаётся вполне приемлемым, поскольку новые данные появляются редко и нет жёстких требований к оперативности.

Итоговая оценка моделей проводилась по совокупности метрик точности (RMSE, MAE, r), полученных на отложенной части реального ряда, и времени обучения, что позволяет учитывать как прогностическую способность, так и практическую применимость метода.

5. РЕЗУЛЬТАТЫ

Расшифровка параметров моделей

- **ARIMA (p, d, q) [целевой ряд]** — ARIMA порядка (p, d, q), где p — порядок авторегрессии (AR), d — порядок интегрирования, q —

порядок скользящего среднего (МА). Целевой ряд — ряд, значения которого восстанавливаются моделью.

- **ARIMAX (p, d, q) [целевой ряд] + [экзогенные переменные]** — модификация ARIMA, включающая экзогенные переменные, используемые в качестве предикторов для повышения точности восстановления целевого ряда.
- **LSTM/GRU [целевой ряд] + [экзогенные переменные], n =[число нейронов], e =[число эпох]** — рекуррентные нейронные сети LSTM и GRU, n — количество нейронов в скрытом слое сети, e — количество эпох обучения.
- **Временные ряды** — $T(30\text{ м})$ — температура воздуха на высоте 30 м; $\text{CO}_2(30\text{ м})$ и $\text{CO}_2(50\text{ м})$ — концентрация углекислого газа на высотах 30 и 50 м соответственно.

5.1. Ряд концентрации CO_2 на высоте 30 м

Анализ временных рядов с использованием модели ARIMA (авторегрессионная интегрированная скользящая средняя) основан на проверке стационарности ряда и выборе подходящих параметров модели. Для этого проводят анализ автокорреляционной функции (АКФ) и частичной автокорреляционной функции (ЧАКФ). АКФ показывает степень корреляции между значениями временного ряда при различных лагах, а ЧАКФ характеризует вклад отдельного лага при фиксировании влияния всех предыдущих лагов [20]. На графиках АКФ и ЧАКФ (рис. 1) светло-синим цветом показаны границы доверительного интервала при уровне значимости 5%, что позволяет визуально выделить лаги со статистически значимой автокорреляцией.

Автокорреляционная функция (АКФ) показывает степень корреляции между значениями временного ряда при различных лагах. Частичная автокорреляционная функция (ЧАКФ) характеризует связь между значениями ряда на данном лаге при фиксировании влияния всех предыдущих лагов. АКФ и ЧАКФ используются для выявления структуры зависимости во временном ряду и выбора порядка авторегрессионных и скользящих средних компонентов модели [21].

В рассматриваемом случае графики АКФ и ЧАКФ (рис. 1) указывают на стационарность ряда, поэтому в модели ARIMA вида (1) были выбраны параметры $p = 3$ и $q = 2$. Параметр $p = 3$ соответствует трем авторегрессионным компонентам $\varphi_1, \varphi_2, \varphi_3$, а параметр $q = 2$ — двум компонентам скользящего среднего θ_1, θ_2 , что отражает наличие статистически значимой автокорреляции на малых лагах. Результат теста Дики-Фуллера ($p\text{-value} = 0.048$) также подтверждает стационарность ряда при уровне

значимости 0.05, поэтому значение $d = 0$ указывает на отсутствие необходимости в дополнительном дифференцировании ряда.

Также была составлена матрица корреляций (табл. 1), которая указала на существенно высокий коэффициент корреляции Пирсона (0.967) между концентрациями углекислого газа на разных высотах. Эти результаты ясно указывают на эндогенный характер показателей концентрации CO_2 на разных высотах, поскольку они проявляют устойчивую и существенную корреляцию, в отличие от предполагаемого экзогенного характера, который исключает такую связь [22].

Таблица 1. Корреляционная матрица метеорологических показателей

	Т (30 м)	$\text{CO}_2(30\text{ м})$	$\text{CO}_2(50\text{ м})$
Т (30 м)	1	0.197	0.127
$\text{CO}_2(30\text{ м})$	0.197	1	0.967
$\text{CO}_2(50\text{ м})$	0.127	0.967	1

В процессе решения задачи восстановления пропусков различной продолжительности в ряде показателя концентрации углекислого газа на высоте 30 м с помощью многомерных моделей были искусственно созданы пропуски продолжительностью от 1 до 60 дней в случайных участках ряда. Используя эти данные, был построен рис. 2, иллюстрирующий зависимости исходных данных и данных, восстановленных с помощью моделей от времени на одном из участков с искусственным пропуском.

Принимая во внимание метрики качества моделей, представленные в табл. 2, можно сделать следующие выводы. Модель ARIMAX (3,0,2) с входными параметрами « $\text{CO}_2(30\text{ м}) + \text{Т}(30\text{ м})$ » демонстрирует наименьшие значения ошибок RMSE и MAE, однако время её восстановления составляет 2.3 с.

Для сравнения: модель ARIMAX (3,0,2) с параметрами « $\text{CO}_2(30\text{ м}) + \text{CO}_2(50\text{ м})$ » показывает практически двукратный выигрыш в скорости, затрачивая на процедуру восстановления 1.13 с.

Обе архитектуры на основе ARIMAX кардинально превосходят по быстродействию нейросетевые подходы: время работы LSTM составляет 251 с, а GRU — 160 с, что в сотни раз медленнее.

5.2. Ряд температуры воздуха на высоте 30 м

Анализируя показатели в табл. 3, мы можем видеть, что модели «ARIMA (3,0,2) Т 30м», «ARIMAX (3,0,2) Т 30м + $\text{CO}_2(50\text{ м})$ » и «ARIMAX (3,0,2) Т 30м + $\text{CO}_2(30\text{ м})$ » не справляются с задачей, поскольку их результаты плохо коррелируют с фактическими значениями температуры воздуха на высоте 30 м. На рис. 3 представлены исходные данные, а также результаты восстановления по всем моделям. Сравнение остальных метрик

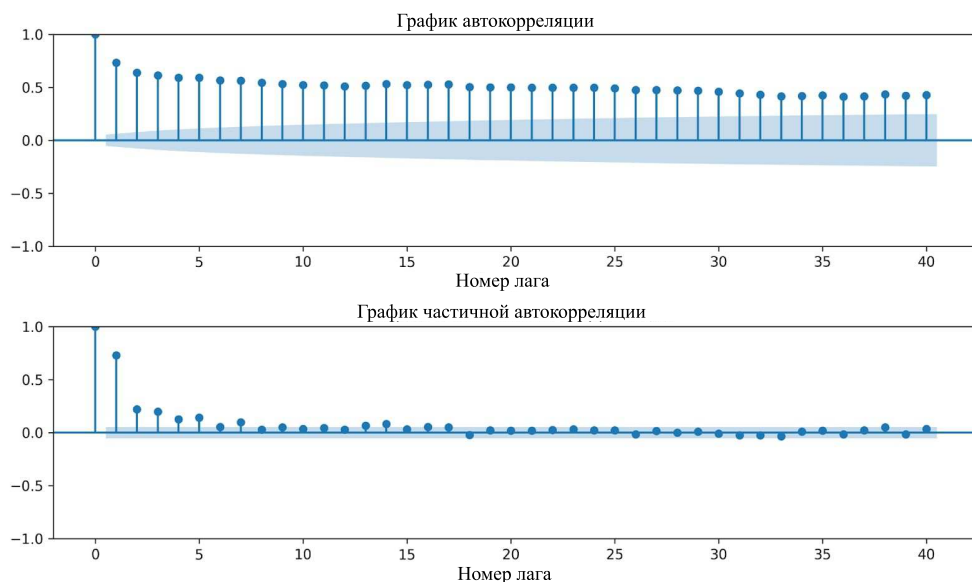


Рис. 1. Автокорреляционная и частичная автокорреляционные функции временного ряда концентрации CO_2 на высоте 30 м

Таблица 2. Анализ метрик различных моделей для восстановления ряда CO_2 на высоте 30 м

Модель	RMSE	MAE	r	t_{learn} , с
ARIMAX (3,0,2) CO_2 30m + T 30m, CO_2 50m	3.35	2.70	0.95	2.3
ARIMAX (3,0,2) CO_2 30m + T 30m	9.86	8.05	0.11	3.3
ARIMAX (3,0,2) CO_2 30m + CO_2 50m	3.38	2.73	0.94	1.13
ARIMA (3,0,2) CO_2 30m	9.49	7.64	-0.09	2.9
LSTM CO_2 30m + CO_2 50m, $n=100$, $e=50$	6.97	5.59	0.81	251
GRU CO_2 30m + CO_2 50m, T 30m, $n=200$, $e=30$	6.89	5.60	0.81	160

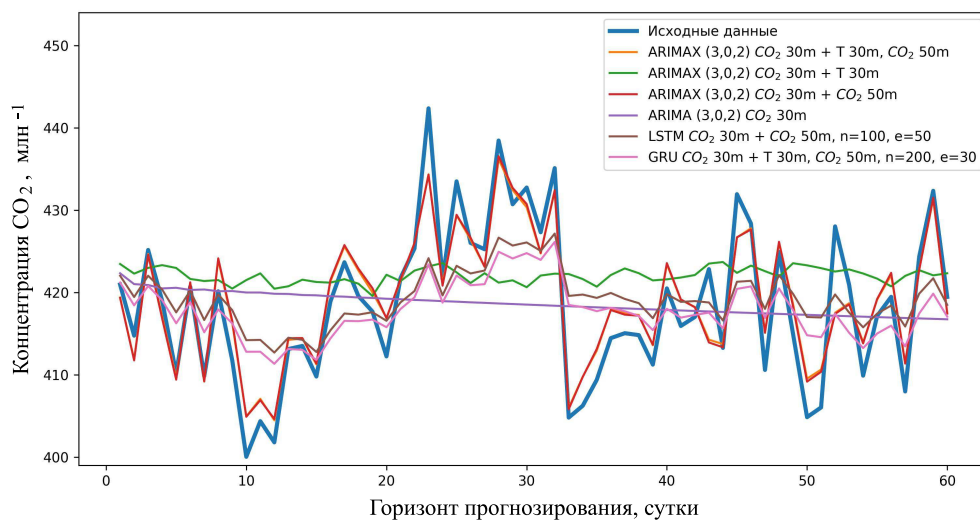


Рис. 2. Все используемые в работе модели восстановления временного ряда концентрации CO_2 на высоте 30 м

указывает на то, что модель «GRU T 30m + CO_2 30m, $n=60$, $e=50$ » превосходит «LSTM T 30m + CO_2 30m, $n=200$, $e=50$ » по ключевым показателям: демонстрируя более низкое значение RMSE (0.77

против 0.86), существенно меньшее время вычислений (192 с против 365 с) и имея лучшее значение корреляции Пирсона ($r = 0.86$ против 0.71). В отличие от моделей ARIMA, GRU обеспечивает опти-

Таблица 3. Анализ метрик различных моделей для восстановления ряда температуры на высоте 30 м

Модель	RMSE	MAE	r	t_{learn} , с
ARIMAX (3,0,2) T 30m + CO ₂ 50m, CO ₂ 30m	1.00	0.79	0.29	3.79
ARIMAX (3,0,2) T 30m + CO ₂ 30m	1.02	0.82	0.12	3.04
ARIMAX (3,0,2) T 30m + CO ₂ 50m	1.04	0.83	0.07	4.01
ARIMA (3,0,2) T 30m	1.06	0.87	-0.06	3.25
LSTM T 30m + CO ₂ 30m, $n=200$, $e=50$	0.86	0.71	0.71	365
GRU T 30m + CO ₂ 30m, $n=60$, $e=50$	0.77	0.63	0.86	192

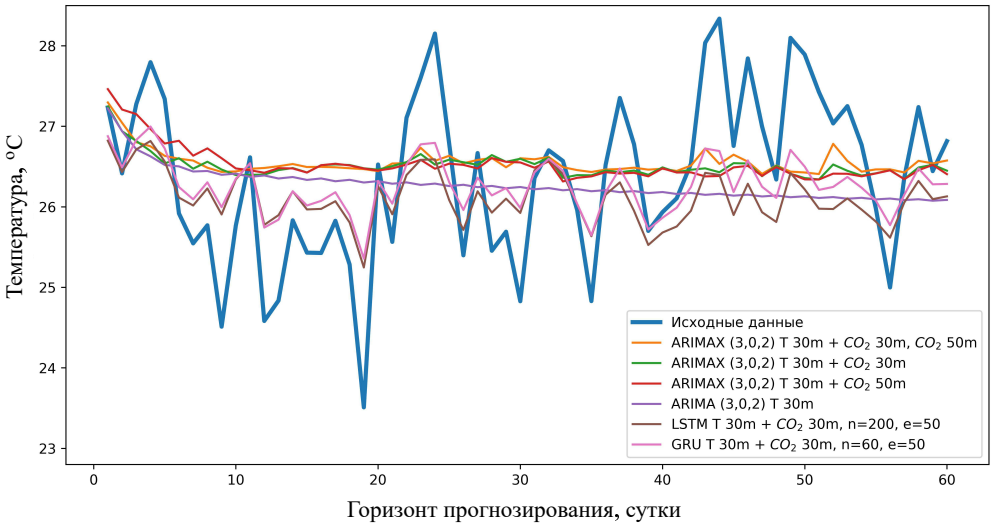


Рис. 3. Все используемые в работе модели восстановления временного ряда температуры на высоте 30 м

мальный баланс между прогностической способностью и вычислительной эффективностью, что делает её предпочтительным выбором для восстановления временного ряда температуры воздуха.

ЗАКЛЮЧЕНИЕ

В результате нашего анализа мы пришли к выводу, что для восстановления ряда концентрации углекислого газа наиболее эффективной оказалась модель семейства ARIMA: «ARIMAX (3,0,2) CO₂ 30m + T 30m, CO₂ 50m», которая использует сильную связь между параметрами CO₂ 30m и CO₂ 50m. В отношении ряда температуры воздуха, для которого не было найдено ни одного ряда с высокой корреляцией, наиболее эффективной оказалась модель на основе рекуррентной нейронной сети «GRU T 30m + CO₂ 30m, $n=60$, $e=50$ ». Это позволяет сделать вывод, что модели семейства ARIMA наиболее подходят для восстановления, ко-

гда имеется другой параметр с высокой корреляцией, в противном случае — предпочтение следует отдать модели нейронной сети LSTM.

С учетом специфики нашего исследовательского вопроса, выбранные методы были адаптированы и успешно реализованы в программной среде Python, являющейся одной из наиболее популярных сред для анализа данных и машинного обучения.

Был выполнен процесс восстановления пропущенных значений во временных рядах концентрации CO₂ и температуры воздуха с применением методов машинного обучения, в частности рекуррентных нейронных сетей LSTM и GRU. Это подтвердило возможность применения этих методов в такого рода задачах.

Исследование выполнено в рамках государственного задания МГУ имени М.В. Ломоносова. Работа Курбатовой Ю.А., Авилова В.К., Куричевой О.А. выполнена в рамках государственного задания Минобрнауки для ИПЭЭ РАН FFER-2024-0025, № 1022033100172-2-1.6.19.

- [1] Авилов В.К., Аleshновский В.С., Безрукова А.В. и др. // Журнал вычислительной математики и математической физики. **61**, N 7. 1113 (2021). (Avilov V.K., Aleshnovskii V.S., Bezrukova A.V. et al. // Computational Mathematics and Mathematical Physics. **67**, N 7. 1106 (2021).)
- [2] Kurbatova J.A., Aleshnovskij V.S., Kuricheva O.A. et al. // IOP Conf. Series: Earth and Environmental Science. **606**. 1 (2020).
- [3] Аleshновский В.С., Безрукова А.В., Авилов В.К. и др. // Вестн. Моск. ун-та. Физ. Астрон. **78**, N 3. 2330101 (2023). (Aleshnovskii V.S., Bezrukova A.V., Avilov V.K. et al. // Moscow Univ. Phys. Bull. **78**, N 3. 73 (2023).)
- [4] Kurbatova J., Tatarinov F., Molchanov A. et al. // Environ. Res. Lett. **8**, N 4. 045028 (2013).
- [5] Тимохина А. В. Динамика концентрации атмосферного диоксида углерода над среднетаежными экосистемами Приенисейской Сибири (по данным измерений на обсерватории ЗОТТО) Красноярск: 2017. С. 49-53.
- [6] Box G., Jenkins G. Time series analysis: Forecasting and control. 1970.
- [7] Пытьев Ю.П., Шишмарев И.А. Теория вероятностей, математическая статистика и элементы теории возможностей для физиков. М., 2010.
- [8] EskandariNasab M., Hamdi S.M., Boubrahimi S.F. et al. // The Astrophysical Journal Supplement Series. **275**. N 1. 6 (2024).
- [9] Newbold P. // Journal of forecasting. **2**. N 1. 23 (1983).
- [10] Faruk D.Ö. // Engineering applications of artificial intelligence. **23**. N 4. 586 (2010).
- [11] Araghinejad S. // Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering. **67**. 85 (2014).
- [12] Wangdi K., Singhasivanon P., Silawan T. et al. // Malaria Journal. **9**. 1 (2010).
- [13] Pektaş A.O., Cigizoglu H.K. // Journal of hydrology. **500**. 21 (2013).
- [14] Alharbi F.R., Csala D. // Inventions. **7**. N 4. 94 (2022).
- [15] Siami-Namini S., Tavakoli N., Namin A.S. // 2018 17th IEEE international conference on machine learning and applications (ICMLA). 1394 (2018).
- [16] Song X., Liu Y., Xue L. et al. // Journal of Petroleum Science and Engineering. **186**. 106682 (2020).
- [17] Yadav H., Thakkar A. // Expert Systems with Applications. **238**. 122333 (2024).
- [18] Cho K., Van Merriënboer B., Gulcehre C. et al. // arXiv preprint arXiv:1406.1078. (2012).
- [19] Yamak P.T., Yujian L., Gadosey P.K. // Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence. 49 (2019).
- [20] Flores J.H.F., Engel P.M., Pinto R.C. // The 2012 International joint conference on neural networks (IJCNN). 1 (2012).
- [21] Abdhussain A., Abed M., Rahi W. // Journal of Al-Rafidain University College For Sciences (Print ISSN: 1681-6870, Online ISSN: 2790-2293) N 1. 554 (2020).
- [22] Priestley M.B., Rao T.S. // Journal of the Royal Statistical Society Series B: Statistical Methodology. **31**. N 1. 140 (1969).

Recovering of Gaps in Multivariate Time Series Using Methods of Mathematical Statistics and Machine Learning

V. S. Aleshnovskii^{1,a}, V. K. Avilov², P. V. Golubtsov¹, Yu. A. Kurbatova², O. A. Kuricheva², D. A. Makeev¹, P. A. Fadeeva¹, A. I. Chulichkov^{1,3}, N. E. Shapkina^{1,b}

¹Faculty of Physics, Lomonosov Moscow State University
Moscow 119991, Russia

²Severtsov Institute of Ecology and Evolution of the Russian Academy of Sciences
Moscow 119071, Russia

³A.M. Obukhov Institute of Atmospheric Physics of the Russian Academy of Sciences
Moscow 119017, Russia

E-mail: ^aaleshnovskii.vs17@physics.msu.ru, ^bneshapkina@mail.ru

The research focuses on multivariate time series of geophysical parameters. During the study, statistical and machine learning methods were employed to recover missing data. The following results were achieved: for the CO₂ concentration series at a height of 30 m, the ARIMAX model demonstrated the lowest RMSE and MAE metrics, taking only 2.3 s for recovery, which is significantly faster compared to the LSTM (251 s) and GRU (160 s) neural network models. Recovery of air-temperature data was performed most effectively using the recurrent neural network GRU. The results indicate that ARIMA family models are more suitable for data recovery when correlated parameters are present, whereas GRU-based neural networks are preferable in other cases. The conclusions can be used to improve data recovery methods in multivariate time series.

PACS: 02.70.Rr, 02.50.Ey

Keywords: multivariate time series, geophysical parameters, data gap recovering, mathematical modeling, carbon dioxide.

Received 09 April 2026.

English version: *Moscow University Physics Bulletin*. 2026. **81**, No. 3. Pp. .

Сведения об авторах

1. Алешновский Валентин Сергеевич — аспирант 3 года обучения; e-mail: aleshnovskii.vs17@physics.msu.ru.
2. Авилов Виталий Константинович — науч. сотрудник; e-mail: vitavilov@gmail.com.
3. Голубцов Петр Викторович — доктор физ.-мат. наук, профессор; e-mail: pgolubtsov@gmail.com.
4. Курбатова Юлия Александровна — канд. биол. наук, доцент, зав. лабораторией;
e-mail: kurbatova.j@gmail.com.
5. Куричева Ольга Алексеевна — канд. биол. наук, науч. сотрудник; e-mail: Olga.Alek.De@gmail.com.
6. Макеев Даниил Александрович — студент 5 курса специалитета; e-mail: makeev05.06daniil2003@mail.ru.
7. Фадеева Пелагея Александровна — студент 6 курса специалитета; e-mail: fadeeva.pa20@physics.msu.ru.
8. Чуличков Алексей Иванович — доктор физ.-мат. наук, зав. кафедрой; e-mail: achulichkov@gmail.com.
9. Шапкина Наталья Евгеньевна — канд. физ.-мат. наук, доцент; тел.: (495) 939-13-51,
e-mail: neshapkina@mail.ru.