ТЕОРЕТИЧЕСКАЯ И МАТЕМАТИЧЕСКАЯ ФИЗИКА

Восстановление пропусков во временных рядах концентрации CO_2 и температуры воздуха методом математической статистики

В. С. Алешновский,[1,*] А. В. Безрукова,[2] В. К. Авилов,[3] В. А. Газарян,[1,4] Ю. А. Курбатова,[3] О. А. Куричева,[3] А. И. Чуличков,[1,5] Н. Е. Шапкина[1,6,†]

[1] *Московский государственный университет имени М. В. Ломоносова, физический факультет. Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2*
[2] *Mars Inc. Россия, 125315, Москва, Ленинградский пр-т, д. 72, к. 1*
[3] *Институт проблем экологии и эволюции имени А. Н. Северцова РАН Россия, 119071, Москва, Ленинский пр-т., д. 33*
[4] *Финансовый университет при Правительстве Российской Федерации Россия, 125993, Москва, Ленинградский пр-т., д. 49*
[5] *Институт физики атмосферы имени А. М. Обухова РАН. Россия, 119017, Москва, Пыжевский пер., д. 3*
[6] *Институт теоретической и прикладной электродинамики РАН. Россия, 125412, Москва, Ижорская ул., д. 13*
(Поступила в редакцию 24.12.2022; после доработки 09.02.2023; принята к публикации 14.02.2023)

Статья посвящена проблеме восстановления пропусков в рядах данных экспериментальных многолетних непрерывных высокочастотных наблюдений за концентрацией диоксида углерода и температурой воздуха. Исследование выполнено на примере результатов наблюдений автоматической эколого-климатической станции, расположенной в тропическом муссонном лесу на территории южного Вьетнама (заповедник Донг Най). Пропуски в рядах наблюдений, как правило, носят случайный характер и обусловлены техническими неисправностями приборной базы. Корректно восстановленные ряды наблюдений позволяют оценить временную изменчивость наблюдаемых параметров на различных временных масштабах. В рамках данного исследования были рассмотрены варианты восстановления непрерывности временных рядов, основанные на методах математической статистики — авторегрессии (ARIMA) и методе линейного прогноза. Приведен сравнительный анализ точности восстановления пропусков различными методами.

PACS: 02.70.Rr, 02.50.Ey УДК: 519.21, 519.25
Ключевые слова: восстановление временных рядов, геофизические данные, корреляция, авторегрессия, диоксид углерода.

DOI: 10.55959/MSU0579-9392.78.2330101

ВВЕДЕНИЕ

Получение качественных данных, анализ которых позволяет оценить современные изменения метеорологических параметров атмосферы, является важнейшей задачей современных климатических и экологических исследований. Благодаря развитию инструментальной базы, средств передачи и хранения данных, вычислительной техники научное сообщество получило возможность использования в исследованиях программного-аппаратных автоматических комплексов для наблюдений за широким спектром абиотических факторов. В настоящее время в рамках исследований взаимодействия наземных экосистем с атмосферой происходит формирование и развитие сети эколого-климатических станций для наблюдений за концентрациями и потоками парниковых газов, а также за метеорологическими величинами на основе высокочастотных автоматических круглогодичных наблюдений (сеть FLUXNET). Непрерывные ряды всех измеряемых параметров являются основой для оценки временной изменчивости процессов, определяющих климаторегулирующие функции экосистем [1, 2]. Однако, несмотря на высокое качество используемых приборов, в рядах данных всегда существуют пропуски, которые носят, как правило, случайный характер и связаны с техническими неисправностями (отсутствие энергоснабжения, выход измерительной аппаратуры из строя, проблемы с калибровками приборов и пр.), что не позволяет обеспечить непрерывность регистрации измеряемых параметров. В этой ситуации восстановление рядов данных осуществляется на основе математического моделирования [3]. Цель настоящего исследования состояла в построении и сравнении нескольких математических моделей, позволяющих восстанавливать ряды динамики. Работа была выполнена на примере данных наблюдений за концентрацией диоксида углерода и температуры воздуха, которые были измерены на эколого–климатической станции, расположенной в тропическом муссонном лесу южного Вьетнама в период с 2011 по 2017 гг. Ранее был проведён анализ временных рядов концентрации углекислого газа на различных высотах над поверхно-

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

стью Земли методами математической статистики, морфологического и вейвлет анализа [1–3]. Полученные результаты будут востребованы в рамках изучения отклика тропических лесов на современные изменения глобального климата.

1. МОДЕЛИ И МЕТОДЫ АВТОРЕГРЕССИИ

1.1. Восстановление пропусков данных во временных рядах концентрации CO_2 методом линейного прогноза

В ходе анализа зависимости между временными рядами концентрации CO_2 на разных высотах над поверхностью Земли была выявлена прямая корреляционная связь [4, 5], в связи с чем для восстановления пропусков данных можно применить метод линейного прогноза [6].

В задаче линейного прогноза необходимо найти линейную функцию $\hat{f}(x_1) = \hat{a_1}x_1 + \hat{a_2}$, удовлетворяющую условию минимизации среднеквадратичной погрешности прогноза:

$$M\left(x_2 - \hat{f}(x_1)\right)^2 = M\left(x_2 - \hat{a_1}x_1 - \hat{a_2}\right)^2 = \min_{\hat{a_1},\hat{a_2}} M\left(x_2 - a_1x_1 - a_2\right)^2, \quad (1)$$

где x_1, x_2 — значения концентрации углекислого газа на двух высотах h_1 и h_2; $\hat{f}(x_1)$ — прогнозируемое с помощью линейной функции значение x_2 на высоте h_2 при известном значении концентрации x_1 на высоте h_1; a_1, a_2 — коэффициенты линейной функции, которые подбираются из условия минимизации выше из произвольного диапазона значений; $M(...)$ — математическое ожидание случайной величины, находящейся в скобках. Как известно [6], решением задачи (1) является

$$\hat{f}(x_1) = M(x_2) + \frac{\mathrm{cov}(x_2, x_1)}{D(x_1)}(x_1 - M(x_1)), \quad (2)$$

а погрешность прогноза равна

$$M\left(x_2 - \hat{f}\left(x_1\right)\right)^2 = D(x_2) - \frac{\left(\mathrm{cov}\left(x_1, x_2\right)\right)^2}{D(x_1)},$$

где $D(x_1)$ и $D(x_2)$ — дисперсии концентраций x_1 и x_2; $\mathrm{cov}\left(x_1, x_2\right)$ — ковариация.

На практике вместо математических ожиданий $M(x_1)$, $M(x_2)$, дисперсий $D(x_1)$, $D(x_2)$ и ковариации $\mathrm{cov}(x_1, x_2)$ используются их оценки в виде средних арифметических значений ряда и средних арифметических квадратов отклонений ряда, полученные из значений ряда за предыдущий отрезок времени (использовались 60 значений ряда).

Заметим, что для получения линейного прогноза нет необходимости знать совместное распределение случайных величин x_2, x_1. В качестве эмпирических оценок математического ожидания, ковариации и дисперсии временных рядов рассматриваются их выборочные значения.

1.2. Восстановление пропусков данных во временных рядах с помощью ARIMA

Одним из прогностических методов восстановления геофизических данных является построение интегрированной модели авторегрессии — скользящего среднего $ARIMA(p, d, q)$ и ее частного случая $ARMA(p, q)$ для стационарного ряда. Модель ARIMA характеризуется тремя параметрами: p — порядок авторегрессии, d — порядок дифференцирования, q — порядок скользящего среднего [8]. Для анализа точности аппроксимации исследуемого временного ряда применяется методология Бокса–Дженкинса, которая заключается в проверке регрессионных остатков на несмещенность, стационарность и неавтокоррелированность [8]. Модель считается адекватной для аппроксимации временного ряда, если все эти свойства выполняются для ряда регрессионных остатков [9].

Стационарный временной ряд со средним значением μ описывается моделью ARMA(p, q), которая имеет следующий вид:

$$x_t = \alpha + \varepsilon_t + \sum_{i=1}^{q} \psi_i \varepsilon_{t-i} + \sum_{i=1}^{p} \theta_i x_{t-i};$$

$$\alpha = \mu\left(1 - \sum_{i=1}^{p} \theta_i\right),$$

где $\theta_1, \ldots, \theta_p, \psi_1, \ldots, \psi_q$ — константы, которые определяются методом наименьших квадратов [9]; ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией. Вводя оператор сдвига L, действующий по правилу $Lx_i = x_{i-1}$, можно записать модель ARMA(p, q) в следующем виде:

$$\Theta(L)x_t = \alpha + \Psi(L)\varepsilon_t,$$

$$\Theta(L) = I - \sum_{i=1}^{p} \theta_i (L)^i; \quad \Psi(L) = I + \sum_{i=1}^{q} \psi_i (L)^i,$$

I — единичный оператор. Нестационарный временной ряд описывается моделью ARIMA(p, d, q). При этом если ряд из разностей его членов, аппроксимирующих дифференцирование ряда порядка d, где d — порядок дифференцирования данного ряда,

$$\nabla^d x_t = (I - L)^d x_t,$$

является стационарным и его удается описать моделью ARMA, тогда соответствующая модель ARIMA(p, d, q) записывается как

$$\Theta(L)\nabla^d x_t = \alpha + \Psi(L)\varepsilon_t.$$

Порядок дифференцирования временного ряда выбирается так, чтобы ряд разностей порядка d был стационарным. Для проверки стационарности временных рядов используется расширенный текст Дики–Фуллера [10].

Чтобы учесть мультипликативную сезонность с периодом S, была использована модель ARIMA$(p, d, q) \times (P, D, Q)_s$, [11], в которой

$$\Theta_p(L)\Theta_P\left(L^S\right)\nabla^d\nabla_S^D x_t = \alpha + \Psi_q(L)\Psi_Q\left(L^S\right)\varepsilon_t.$$

Согласно методологии Бокса–Дженкинса для оценки параметров p и q модели ARIMA используется анализ автокорреляционной и частичной автокорреляционной функций [12]. Значения параметров P и Q сезонной компоненты модели ARIMA выбираются также на основе анализа автокорреляционной и частичной автокорреляционной функций, а параметр D, аналогичный параметру d, также находится путем взятия уже сезонных разностей так, чтобы ряд был стационарным [12]. При наличии сезонной компоненты у временного ряда на графиках этих функций будут наблюдаться характерные максимумы в лагах, соответствующих периоду S сезонной компоненты.

Автокорреляционная функция ACF_τ с лагом автокорреляции τ для временного ряда x вычисляется по формуле [13]:

$$\mathrm{ACF}_\tau = \frac{\sum_{i=1}^{T-\tau}(x_i-\bar{x})(x_{i+\tau}-\bar{x})}{\sum_{i=1}^{T}(x_i-\bar{x})^2}; \quad \bar{x}=\frac{1}{T}\sum_{i=1}^{T}x_i,$$

где T — число отсчетов ряда, используемых для оценки автокорреляционной функции.

Частичная автокорреляционная функция PACF_τ с лагом автокорреляции τ для стационарного временного ряда x вычисляется следующим образом [14]:

$$\mathrm{PACF}_\tau = \begin{cases} M\left[x_{t+1}x_t\right], & \tau = 1; \\ M\left[\left(x_{t+\tau}-x_{t+\tau}^{\tau-1}\right)\left(x_t-x_t^{\tau-1}\right)\right], & \tau \geqslant 2 \end{cases};$$

$$x_t^{\tau-1} = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{\tau-1}x_{t+\tau-1};$$

$$x_{t+\tau}^{\tau-1} = \beta_1 x_{t+\tau-1} + \beta_2 x_{t+\tau-2} + \cdots + \beta_{\tau-1}x_{t+1},$$

где $\beta_1, \cdots, \beta_{\tau-1}$ — коэффициенты линейной регрессии.

Выбор начальных приближений параметров p, q, P и Q осуществляется из следующих соображений [15]:

1. В модели ARIMA$(p,d,0)$ автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а частичная автокорреляционная функция значимо отличается от нуля при лагах, не больших p.

2. В модели ARIMA$(0,d,q)$ частичная автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а автокорреляционная функция значимо отличается от нуля при лагах, не больших q.

3. Начальное приближение для Q задается как номер последнего сезонного лага, при котором автокорреляция значима.

4. Начальное приближение для P задается как номер последнего сезонного лага, при котором частичная автокорреляция значима.

Значимость коэффициента корреляции или автокорреляции определяется в математической статистике следующим образом: выдвигается нулевая гипотеза о незначимости коэффициента корреляции (равенстве нулю) против альтернативы о его значимости на заданном уровне значимости (мы выбрали 0.01). Строится статистика критерия Стьюдента и критическая область. Если статистика критерия попадает в критическую область, гипотеза о незначимости коэффициента корреляции отвергается, в противном случае принимается.

После задания начальных приближений конечный выбор значений четырех параметров p, q, P и Q, как правило, происходит перебором: то есть перебираются всевозможные наборы значений и ищется модель, у которой получилось минимальное значение критерия Акаике [16]. Оптимальной по критерию Акаике будет модель, у которой значение этого критерия наименьшее из всех возможных.

После оптимизации параметров модели проводится анализ остатков, включающий проверку трех условий: несмещенности, стационарности и неавтокоррелированности [17]. При выполнении всех этих условий модель признается адекватной для аппроксимации анализируемого временного ряда, согласно критерию Стьюдента [18].

2. АЛГОРИТМ ВОССТАНОВЛЕНИЯ ПРОПУСКОВ ДАННЫХ ВО ВРЕМЕННЫХ РЯДАХ

В этом разделе описаны необходимые этапы предобработки и восстановления временных рядов [19].

Этап 1 — проверка измерений (валидация) и обнаружение ошибок и пропусков данных. Под ошибками измерений подразумеваются выбросы в измерениях, различные аппаратные ошибки, например, последовательности нулей, значения измерений, соответствующие сбоям датчиков и т. д. Алгоритм обнаружения пропусков проверяет измерения на наличие пропущенных значений путем сравнения временных дискретов $\Delta_i = t_i - t_{i-1}$ каждого i-го измерения с задаваемой величиной дискретизации. Если Δ_i превышает величину дискретизации, то отмечается пропуск между измерениями.

Этап 2 — интерполяция сигналов (временных рядов) на единую временную сетку. На этом этапе выполняется интерполяция измерений с разной частотой дискретизации на временную сетку с постоянным шагом. Шаг сетки выбирается исходя из априорных представлений о гладкости зависимости концентрации CO_2 от времени.

Этап 3 — восстановление пропусков. Оцениваются отсутствующие значения временных рядов, длительность пропущенного фрагмента может варьироваться от одного шага ряда до нескольких сотен.

Этап 4 — запись результатов в базу данных. Восстановленные и синхронизированные значения временного ряда записываются в базу данных, места

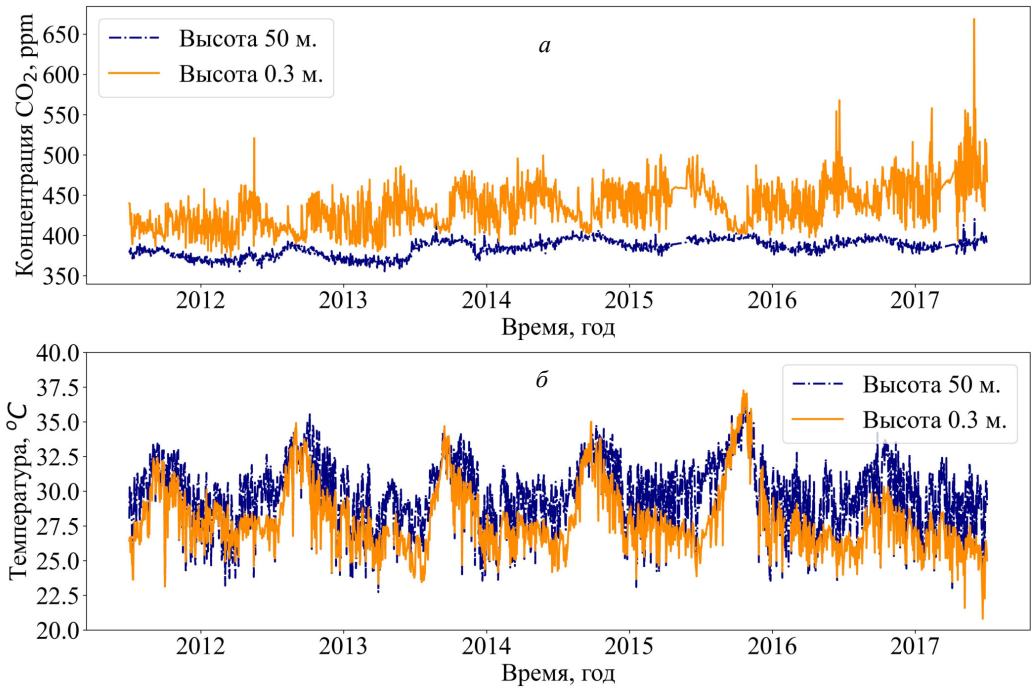

Рис. 1. Временная изменчивость средних значений концентрации CO_2 в дневное время суток ($а$) и среднесуточной температуры воздуха ($б$) на высотах 0.3 и 50 м. Заповедник Донг Най (Вьетнам), период измерений 2011–2017 гг.

восстановленных значений помечаются специальным флагом в базе данных.

После этапа интерполяции сигналов формируются измерения с постоянной частотой дискретизации, описываемые временными рядами $x = \{x(t), t = T\{0 \dots N\}\}$, где T – множество отсчетов времени, $x \in R$. Значения в некоторых отсчетах времени t отсутствуют (пропуски) [20]. Необходимо найти оценку $\hat{f}(t)$ значений сигнала $x(t)$ в местах пропусков. Априорно параметры модели сигнала неизвестны, имеются только исторические записи сигналов (временных рядов) [21].

2.1. Сравнение алгоритмов восстановления пропусков

Для удобства введем обозначения моделей следующим образом:

Модель-1 — восстановление с помощью метода линейного прогноза,

Модель-2 — восстановление с помощью ARIMA.

Две описанные модели были применены для восстановления пропущенных данных, полученных при измерении температуры и концентрации CO_2 на экспериментальной вышке на станции *AsiaFlux* во Вьетнаме. На вышке была установлена система из 8 датчиков температуры и концентрации CO_2 на 8 различных высотах (50, 28, 19, 10, 5, 2, 1, 0.3 м) над уровнем земли. Дискретизация начальных сигналов составляла 1 мин.

Для моделирования было выбрано два сигнала: первый с датчика на вершине измерительной вышки на высоте 50 м над уровнем земли, второй с датчика вблизи земли на высоте 0.3 м, а также два сигнала на промежуточных высотах, имеющих наиболее сильную корреляционную зависимость с первыми двумя сигналами. Примеры сигналов приведены на рис. 1, *а, б*.

Для получения точностных характеристик восстановления пропусков в каждом из сигналов были созданы случайные искусственные пропуски длительностью от недели до двух месяцев. Для всех сигналов пропуск каждой величины генерировался несколько раз в случайные моменты времени и проводилось восстановление пропущенных данных.

Для оценивания точности восстановления была выбрана квадратичная ошибка (MSE), она рассчитывалась следующим образом:

$$MSE = \frac{1}{N+1} \sum_{i=0}^{N} \left(\hat{f}_i(t) - x_i(t) \right)^2,$$

где t — момент времени из множества отсчетов времени $T\{0 \dots N\}$, $\hat{f}(t)$ — прогноз значения показателя в момент времени t, $x(t)$ — известное значение прогнозируемого показателя в момент времени t.

Ниже на рис. 2, *а, б* приведены графики зависимости значений погрешности от длительности восстанавливаемого фрагмента ряда для обоих методов. Как видно из графиков, на высоте 0.3 м лучшие результаты восстановления пропусков данных показала Модель–2, поскольку погрешность восстановления оказалась наименьшей во всех случаях вне зависимости от длины восстанавливаемого пропуска. Предполагается, что это связано с тем, что на малых высотах влияние взаимодействия атмо-

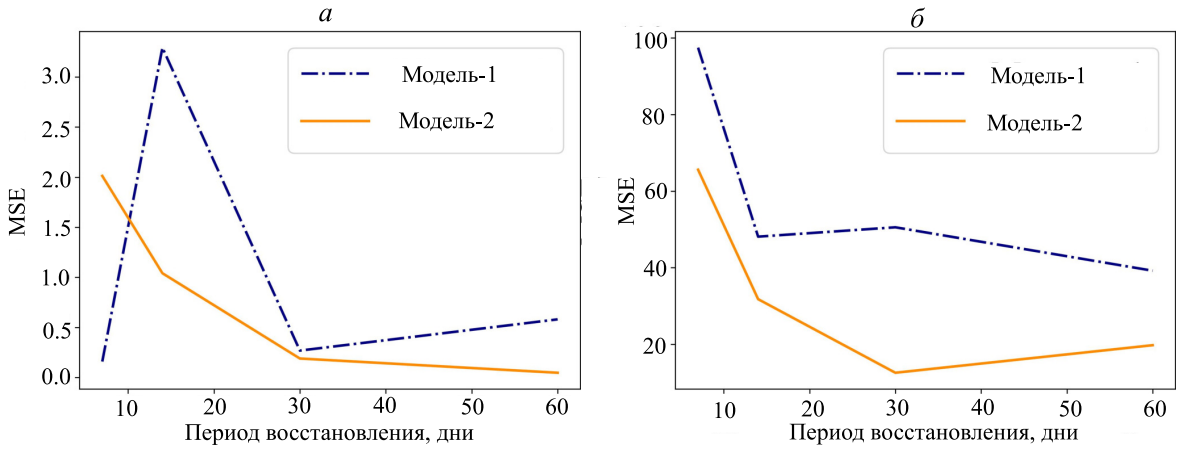

Рис. 2. Зависимость ошибки MSE от длительности восстанавливаемого фрагмента ряда динамики концентрации углекислого газа на высотах 50 м (*а*) и 0.3 м (*б*), где MSE измеряется в ppm^2

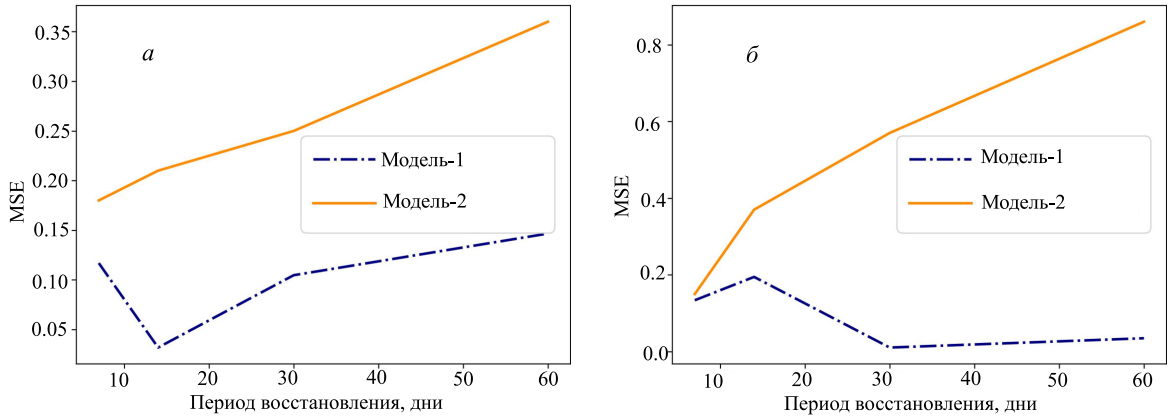

Рис. 3. Зависимость ошибки MSE от длительности восстанавливаемого фрагмента динамики температуры на высотах 50 м (*а*) и 0.3 м (*б*), где MSE измеряется в $\text{C}^{\circ\,2}$

сферы и почвы является наиболее тесным (значительным) [7], чем на больших.

При анализе зависимости величины погрешности от длительности восстанавливаемого фрагмента временных рядов температуры воздуха получены иные результаты. На графиках рис. 2, *а*, *б* можно заметить, что значения концентрации меняются с высотой, что, скорее всего, связано с более интенсивным перемешиванием слоев газа на больших высотах, в отличие от значений температуры, которые мало меняются в зависимости от высоты для рассматриваемых данных. Характер зависимости значений погрешности от длительности восстанавливаемого фрагмента для ряда динамики температуры имеет вид, представленный на рис. 3, *а*, *б*:

На рис. 3, *а*, *б* заметно, что при восстановлении пропущенных данных ряда температуры лучшие результаты получены с помощью Модели-1 вне зависимости от высоты и длины восстанавливаемого промежутка, причем с увеличением длины пропуска значение погрешности возрастает. Это объясняется тем, что Модель-2 восстанавливает данные на основании наблюдений, полученных за предыдущий период период времени, в то время как Модель-1 использует для прогнозирования значения временного ряда на соседней высоте. Так как при прогнозировании температуры значения на разных высотах не сильно отличаются, то результат в данном методе получился лучше. Погрешность оценивания при восстановлении пропущенных значений ряда температуры с помощью Модели-2 достаточно мала, что позволяет применять данную модель при отсутствии измерений на соседних высотах.

В табл. 1, 2 приведены результаты восстановления данных с помощью Модели-1 (М-1) и Модели-2 (М-2) для двух разных высот: 50 м и 0.3 м. Восстановление проводилось для промежутков в 7, 14, 30 и 60 дней.

Анализ моделей показал, что на высоте 50 м Модель-2 восстанавливает концентрацию CO_2 в среднем лучше, чем Модель-1, но на высоте 0.3 м точность восстановления данных обоими методами не очень высокая. При восстановлении пропусков рядов температуры Модель-1 показала более точные

Таблица 1. Точность восстановления рядов динамики температуры воздуха и концентрации углекислого газа на высоте 50 м по двум моделям на основе оценки квадратичной ошибки (MSE ($^\circ C^2$))

Модель	М-1	М-2	М-1	М-2	М-1	М-2	М-1	М-2
Период восстановления	7 дн	7 дн	14 дн	14 дн	30 дн	30 дн	60 дн	60 дн
MSE CO_2	0.16	**0.15**	3,29	**1.04**	0.27	**0.19**	0.58	**0.05**
MSE темп.	**0.12**	0.18	**0.03**	0.21	**0.11**	0.25	**0.15**	0.36

Таблица 2. Точность восстановления рядов динамики температуры воздуха и концентрации углекислого газа на высоте 0.3 м по двум моделям на основе оценки квадратичной ошибки (MSE (ppm^2))

Модель	М-1	М-2	М-1	М-2	М-1	М-2	М-1	М-2
Период восстановления	7 дн	7 дн	14 дн	14 дн	30 дн	30 дн	60 дн	60 дн
MSE CO_2	97.46	**65.58**	48.09	**31.75**	50.54	**12.54**	39.19	**19.74**
MSE темп.	**0.13**	0.15	**0.19**	0.37	**0.01**	0.57	**0.03**	0.86

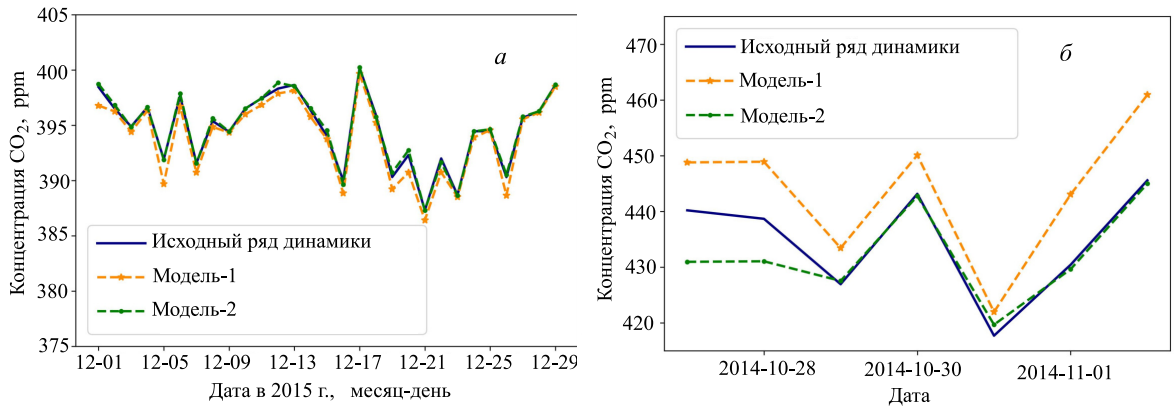

Рис. 4. Динамика измеренных и рассчитанных по моделям средних дневных значений концентрации CO_2 на высоте 50 м для восстанавливаемого фрагмента длительностью 30 дней (*а*) и на высоте 0.3 м для фрагмента длительностью 7 дней (*б*)

результаты, чем Модель-2 для обоих высот вне зависимости от длины промежутка. Следует также отметить, что применение обоих методов восстановления данных приводит к существенно лучшим результатам, чем используемая ранее на станции линейная интерполяция [8].

На рис. 4, *а*, *б* и рис. 5, *а*, *б* ниже представлены результаты восстановления значений концентрации углекислого газа и температуры воздуха с помощью различных моделей на разных высотах.

Таким образом, опираясь на значения погрешностей (табл. 1, 2), можно сделать вывод, что применение Модели-1 приводит к лучшим результатам при восстановлении пропущенных данных в рядах среднесуточной температуры воздуха, а Модели-2 — при восстановлении пропущенных измерений концентрации CO_2. В целом для обеих моделей погрешность восстановления меньше для высоты 50 м, чем для высоты 0.3 м.

ЗАКЛЮЧЕНИЕ

На основании результатов применения методов восстановления пропусков в рядах наблюдений на основе метода линейного прогноза и интегрированной модели авторегрессии показано, что оба метода в целом адекватно описывают временную изменчивость концентрации диоксида углерода и температуры воздуха на различных высотах в пологе тропического муссонного леса в широком диапазоне погодных условий. Такой вывод подтверждается достаточно высокой точностью восстановления пропу-

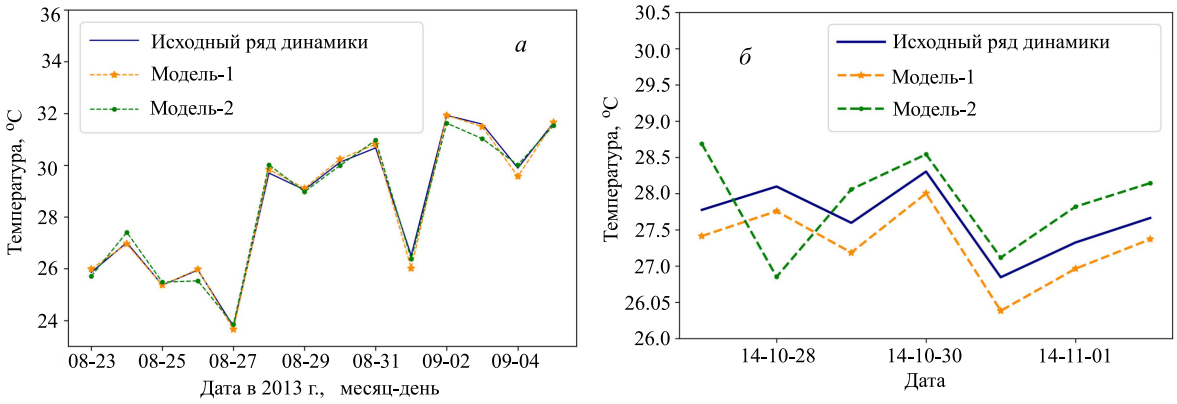

Рис. 5. Динамика измеренных и рассчитанных по моделям средних суточных значений температуры воздуха на высоте 50 м для восстанавливаемого фрагмента длительностью в 14 дней (*а*) и на высоте 0.3 м для фрагмента длительностью 7 дней (*б*)

щенных данных во временных рядах исследуемых метеорологических показателей.

На примере многолетних экспериментальных данных наблюдений показано, что для восстановления пропусков в наблюдениях за температурой воздуха лучше использовать модель на основе метода линейного прогноза, однако при отсутствии необходимых для линейного прогноза данных на близких высотах допустимо применение модели авторегрессии. Для восстановления данных концентрации CO_2, как правило, лучше применять интегрированную модель авторегрессии.

Результаты исследования планируется использовать в рамках работ по изучению временной изменчивости метеорологических показателей и концентрации CO_2 в пологе тропического муссонного леса в условиях современных изменений глобального климата.

Работа выполнена при финансовой поддержке Российского научного фонда (грант 21-14-00209).

[1] *Авилов В.К., Алешновский В.С., Безрукова А.В.* и др. // Ж. вычисл. матем. и матем. физ. **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] *Kurbatova J., Tatarinov F., Molchanov A.* et al. // Environ. Res. Lett. **8**, N 4. 045028. (2013).

[4] *Тимохина А. В.* Динамика концентрации атмосферного диоксида углерода над среднетаежными экосистемами Приенисейской Сибири (по данным измерений на обсерватории «ZOTTO») Красноярск: 2017. — С. 49-53.

[5] *Anggraeni W., Vinarti R.A., Kurniawati Y.D.* // Procedia Computer Science. **72**. 630. (2015).

[6] *Пытьев Ю. П., Шишмарев И. А.* // Теория вероятностей, математическая статистика и элементы теории возможностей для физиков. М., 2010.

[7] *Box G., Jenkins G.* // Time series analysis: Forecasting and control. 1970.

[8] *Hocke K., Kampfer N.* // Atmos. Chem. Phys. **9**, N 12. 4197. (2009).

[9] *Akaike H.* // IEEE Transactions on Automatic Control. **19**, N 6. 716. (1974).

[10] *Gluhovsky A., Ernest A.* // Journal of Applied Meteorology and Climatology. **46**, N 7. 1125. (2007).

[11] *Gallop C., Tseand C., Zhao J.* // i-Manager's Journal on Civil Engineering. **1**, N 4. 9. (2011).

[12] *Dickey D. A., Fuller W.* // Journal of the American Statistical. **74**, N 366a. 427. (1979).

[13] *Alsharif M.H., Kim J., Kim J.H.* // Energies. **10**, N 5. 587. (2017).

[14] *Bai, S., Kolter, J.Z., Koltun, V.* // arXiv preprint arXiv:1803.01271. (2018).

[15] *Brocardo M. L., Traore I., Wonggang I., Obaidat M. S.* // International Journal of Communication Systems. **30**, N 12. e3259. (2017).

[16] *Makridakis S., Spiliotis E., Assimakopoulos V.* // PLoS ONE. **13**, N 3. e0194889. (2018).

[17] *Fattah, J., Ezzine, L., Aman, Z.* et al. // International Journal of Engineering Business Management. **10**. 1847979018808673. (2018).

[18] *Tong M., Duan H., He L.* // Environmental Science and Pollution Research. **28**, N 24. 31370. (2021).

[19] *Leerbeck, K., Bacher, P., Junker, R. G.* et al. // Applied Energy. (2020). **277**. 115527.

[20] *Duchon C., Hale R.* // Time Series Analysis in Meteorology and Climatology: An Introduction John Wiley & Sons. (2012).

[21] *Rahman A., Hasan M. M.* // Open Journal of Statistics. **7**, N 4. 560. (2017).

Gap Recovery in the Time Series of CO_2 Concentration and Air Temperature Using Methods of Mathematical Statistics

V. S. Aleshnovskii[1,a], **A. V. Bezrukova**[2], **V. K. Avilov**[3], **V. A. Gazaryan**[1,4], **Yu. A. Kurbatova**[3], **O. A. Kuricheva**[3], **A. I. Chulichkov**[1,5], **N. E. Shapkina**[1,6,b]

[1] *Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia*
[2] *Mars Inc., Moscow 125315, Russia*
[3] *Severtsov Institute of Ecology and Evolution of the Russian Academy of Sciences, Moscow 119071, Russia*
[4] *Financial University under the Government of the Russian Federation, Moscow 125993, Russia*
[5] *Obukhov Institute of Atmospheric Physics of the Russian Academy of Sciences, Moscow 119017, Russia*
[6] *Institute of Theoretical and Applied Electrodynamics, RAS, Moscow 125412, Russia*
E-mail: [a] *aleshnovskii.vs17@physics.msu.ru*, [b] *neshapkina@mail.ru*

The article is dedicated to the problem of recovering gaps in data series of experimental long-term continuous high-frequency observations of carbon dioxide concentration and air temperature. The study was conducted using the observation results from an automatic eco-climatic station located in a tropical monsoon forest in southern Vietnam (Dong Nai biosphere reserve). Gaps in observation series are, as a rule, random and caused by technical malfunctions of the instrumentation. Accurately recovered observation series allow for the assessment of the temporal variability of observed parameters on different time scales. In the scope of this study, options for recovering the continuity of time series based on mathematical statistics methods—autoregression (ARIMA) and the linear prediction method—have been considered. A comparative analysis of the accuracy of gap recovering using different methods is provided.

PACS: 02.70.Rr, 02.50.Ey
Keywords: time series recovering, geophysical data, correlation, autoregression, carbon dioxide.
Received 24 December 2022.
English version: *Moscow University Physics Bulletin.* 2023. **78**, No. 3. Pp. 324–331.

Сведения об авторах

1. Алешновский Валентин Сергеевич — студент 2 курса магистратуры; e-mail: aleshnovskii.vs17@physics.msu.ru.
2. Безрукова Александра Владимировна — Leadership Program Specialist; e-mail: aleksandra_bezrukova@mail.ru.
3. Авилов Виталий Константинович — науч. сотрудник; e-mail: vitavilov@gmail.com.
4. Газарян Варвара Арамовна — канд. физ.-мат. наук, доцент; тел.: (495) 939-41-78, e-mail: varvaragazaryan@yandex.ru.
5. Курбатова Юлия Александровна — канд. биол. наук, доцент, зав. лабораторией; e-mail: kurbatova.j@gmail.com.
6. Куричева Ольга Алексеевна — канд. биол. наук, науч. сотрудник; e-mail: Olga.Alek.De@gmail.com.
7. Чуличков Алексей Иванович — доктор физ.-мат. наук, зав. кафедрой; тел.: (495) 939-41-78, e-mail: achulichkov@gmail.com.
8. Шапкина Наталья Евгеньевна — канд. физ.-мат. наук, доцент; тел.: (495) 939-13-51, e-mail: neshapkina@mail.ru.