Статистические методы и методы машинного обучения в расчётах уровней воды рек по данным спутниковых альтиметрических измерений

Н. К. Семенова [1, 2], Е. А. Захарова [1, 3], И. Н. Крыленко [1, 2], А. А. Сазонов [1, 2]

[1] *Институт водных проблем РАН, Москва, 119333, Россия*
E-mail: snkone132@mail.ru
[2] *Московский государственный университет имени М. В. Ломоносова Москва, 119991, Россия*
[3] *Спутниковые наблюдения в обучении и в практических приложениях Тулуза, 31400, Франция*

Использование данных спутниковой альтиметрии для мониторинга уровенного режима рек арктических регионов ограничено из-за влияния сложной морфометрии речных долин и ледового покрова на измерения альтиметрических радаров. Построение временных серий уровней воды в реках состоит из двух основных этапов: 1) точной географической выборки спутниковых измерений над руслом реки, 2) вычисления среднего за данную дату уровня после фильтрации выбросов. Данная работа основана на измерениях европейских альтиметрических спутников Sentintel-3A и Sentinel-3B. Предложен метод определения аберрантных значений альтиметрических измерений (выбросов) над широкопойменным участком р. Колымы, который позволил улучшить точность расчёта спутниковых временных серий уровня воды на 0,04–1,59 м (или 4–85 %) по сравнению с широко используемым стандартным статистическим методом фильтрации альтиметрических измерений. Разработанный метод основан на комбинировании трёх алгоритмов различной сложности: статистического (расстояние Махаланобиса), кластеризационного (*англ.* Density-Based Spatial Clustering of Applications with Noise — DBSCAN) и метода машинного обучения («изолирующий лес»). При комбинированном подходе выбросами считались значения, классифицированные таковыми как минимум двумя алгоритмами. Подобный подход позволил уменьшить влияние потенциальных индивидуальных недостатков каждого из трёх методов.

Ключевые слова: спутниковая альтиметрия, уровни воды, арктические реки, определение выборочных выбросов, методы машинного обучения

Одобрена к печати: 08.12.2023
DOI: 10.21046/2070-7401-2024-21-1-76-87

Введение

В последние два десятилетия активно развивается применение альтиметрических спутников, изначально предназначенных для исследования водной поверхности океана, для наблюдений за уровнями воды рек (Abdalla et al., 2021). Основная проблема обработки сигнала альтиметрического радара над реками и причины низкой точности расчётных речных уровней воды (0,1–2,5 м) по сравнению с уровнями морей и океанов (до 0,04 м) связаны с пространственной неравномерностью шероховатости отражающей поверхности в наземной проекции радара. В результате форма волны регистрируемого альтиметром отражённого сигнала имеет многопиковый характер, где пики соответствуют областям с низкой шероховатостью. Если такая область лежит вне оси надира инструмента, большинство современных алгоритмов расчёта высоты поверхности (т. е. ретрекингов) будут определять расстояние до этой области, а не до русла реки в надире. В этом случае может наблюдаться как занижение расчётных уровней (если поверхность находится на одном уровне с основным руслом, как, например, малый речной рукав), так и завышение, если объект находится на пойме (старицы) (Biancamaria et al., 2018). Дополнительную сложность создают песчаные осерёдки, влажные пески которых отражают больше энергии радара, чем взволнованная водная поверхность реки (Maillard et al., 2015). В этом случае расчётный уровень воды в межень может быть завы-

шен. Предсказать форму волны отражённого сигнала в каждом конкретном месте пересечения трека радара с руслом реки (на так называемой виртуальной станции) невозможно. В результате различные ретрекеры, как официальные, так и разработанные научными коллективами, показывают эффективность, сильно варьирующую от места к месту и в зависимости от фазы водного режима.

У альтиметров последнего поколения с синтезированной апертурой (*англ*. Synthetic Aperture Radar Interferometer Radar Altimeter (SIRAL) и Synthetic Aperture Radar Altimeter (SRAL)), установленных на спутниках CryoSat-2, Sentinel-3A/B и Sentinel-6, площадь отражающей поверхности значительно уменьшена вдоль направления движения спутника (вдоль трека). Это позволило частично решить вопрос о влиянии водных объектов, расположенных вне надира радара. Тем не менее в плоскости, перпендикулярной треку, площадь иллюминированной земной поверхности остаётся значительной: ±1–3 км в обе стороны от надира (Rémy et al., 2012) в зависимости от шероховатости водной поверхности в основном русле. Загрязнение эха радара сигналом, отражённым от рукавов, стариц и осерёдков, может также приводить к ошибкам в расчётах уровней воды в основном русле. Для более точного позиционирования принимающего окна радара последние европейские альтиметры Sentinel-3A/B и Sentinel-6 работают в режиме открытого контура (*англ*. open loop) и используют априорную базу данных о высоте местности из цифровой модели рельефа (ЦМР), хранящейся на борту. Такой режим позволяет корректно и быстро позиционировать окно приёма отражённого сигнала радара в сложных топографических условиях, благодаря чему можно значительно расширить область применения альтиметрии, в частности для рек со средними (Колыма) и высокими (Маккензи, Лена) надпойменными террасами.

Несмотря на значительный прогресс в технологиях и алгоритмах, решение проблемы точности расчётов спутниковых уровней воды в реках далеко от завершения. Были разработаны множественные подходы автоматизированной выборки измерений радара, а также расчётов уровня из полученной выборки. Наиболее распространённый подход — использование масок водных поверхностей (в растровом или векторном формате) с географической привязкой и вычисление медианных значений уровня за каждый пролёт спутника (Zakharova et al., 2020). Заключительным этапом обработки становится редактирование полученных временных серий на предмет выявления и выбраковки аберрантных измерений уровней. Традиционно используется статистический подход, признающий выбросами значения, в два или три раза превышающие стандартное выборочное отклонение (Biancamaria et al., 2018; Maillard et al., 2015). К нему может добавляться фильтрация по глобальному среднему, по текущему среднему или по скользящему среднему (ATBD..., 2022). Есть примеры использования фильтра Кальмана (Schwatke et al., 2015). В представляемой работе было решено разработать и протестировать новый подход к выбраковке аберрантных измерений (выбросов), комбинирующий простые статистические методы с более продвинутыми — кластеризационным и методом машинного обучения.

Исходные данные

Измерения альтиметрических спутников

В качестве входных данных использовались данные европейских альтиметрических спутников Sentinel-3A (S3A) и Sentinel-3B (S3B), запущенных соответственно в 2016 и 2018 гг. Спутники летают по солнечно-синхронной орбите со средней высотой 815 км и имеют 27-дневный цикл. Миссии оснащены двухчастотным альтиметрическим радаром SRAL, работающим в частотных диапазонах Ku и C. Для расчётов уровней традиционно используются измерения радара в диапазоне Ku. В данной работе было отдано предпочтение алгоритму обработки отражённого сигнала SAMOSA, который в предыдущем исследовании на том же объекте показал наилучшие результаты в зимний период (Захарова и др., 2023). Измерения радара проводятся с частотой 20 Гц, что соответствует 300 м между соседними измерениями. Геофизические коррекции, такие как модельные коррекции влажной и сухой тропосферы,

ионосферная коррекция, коррекции на приливы земной коры, поставляемые в спутниковом продукте с частотой 1 Гц, были линейно проинтерполированы к координатам измерений радара с частотой 20 Гц. Уровень воды вычислялся по следующей формуле:

$$\mathrm{WL} = \mathrm{Alt} - R + \mathrm{H}_{\mathrm{wetTr}} + \mathrm{H}_{\mathrm{dryTr}} + \mathrm{H}_{\mathrm{iono}} + \mathrm{H}_{\mathrm{polTide}} + \mathrm{H}_{\mathrm{solidTide}} - \mathrm{H}_{\mathrm{geoid}},$$

где WL — уровень воды; Alt — высота спутниковой орбиты; R — расстояние от антенны до водной поверхности; $\mathrm{H}_{\mathrm{wetTr}}$ — коррекция на замедление сигнала во влажной атмосфере; $\mathrm{H}_{\mathrm{dryTr}}$ — коррекция на замедление сигнала в сухой атмосфере; $\mathrm{H}_{\mathrm{iono}}$ — ионосферная коррекция; $\mathrm{H}_{\mathrm{solidTide}}$ и $\mathrm{H}_{\mathrm{polTide}}$ — высотные коррекции за счёт деформаций и приливов земной коры; $\mathrm{H}_{\mathrm{geoid}}$ — отклонение геоида от эллипсоида.

Натурные данные

Для проверки эффективности алгоритмов фильтрации использовались данные по уровням воды гидрометрического поста с. Колымское за период 2016–2021 гг. Натурные данные за зиму 2018 г. были исключены из рассмотрения ввиду сомнительных значений уровня воды, превышающих на 1 м типичный уровень в этот сезон. Правомерность исключения была проверена по расходам воды вышерасположенного гидрологического поста г. Среднеколымска, на котором зимние расходы воды 2018 г. не отличались от типичных значений за предыдущие и последующие зимы. Больших речных притоков, могущих объяснить дополнительное поступление воды между постами Среднеколымское и Колымское, тем более в зимний период, не существует.

Методы

Описание алгоритмов фильтрации

В работе были использованы четыре алгоритма, позволяющих определить выбросы в выборках данных. Два алгоритма относятся к стандартным статистическим методам: правило трёх сигм (3σ) и расстояние Махаланобиса. Третий выбранный алгоритм — плотностной алгоритм пространственной кластеризации с присутствием шума (*англ.* Density-Based Spatial Clustering of Applications with Noise — DBSCAN) — основан на кластеризации данных. Более сложный четвёртый алгоритм — «изолирующий лес» (*англ.* Isolation Forest — IF) — относится к классу алгоритмов машинного обучения.

Статистические методы

Правило трёх сигм — это классический метод, основанный на неравенстве Чебышева

$$P\left(\left|\xi - E\xi\right| \geq a\right) \leq \frac{\sigma^2}{a^2},$$

где a — константа, $a > 0$; σ — стандартное отклонение (сигма); ξ — произвольная случайная величина с конечным математическим ожиданием и дисперсией; Eξ — математическое ожидание ξ и $P(...)$ означает вероятность указанного события.

Правило подразумевает, что для любой случайной величины с конечной дисперсией вероятность отклонения от своего математического ожидания более чем на три стандартных отклонения не превосходит 1/9 (в случае нормального распределения эта вероятность не превосходит 0,28 %). В практическом применении алгоритма 3σ все значения в выборке, отклоняющиеся от своего математического ожидания более чем на 3σ, считаются выбросами.

Расстояние Махаланобиса часто используется в нахождении выбросов в многомерных данных. Расстояние Махаланобиса — это расстояние от каждого наблюдения до центра облака точек. В отличие от евклидова расстояния, где данные считаются нормально распределёнными, расстояние Махаланобиса учитывает корреляцию между признаками и может ра-

ботать и в случае, когда данные сильно скоррелированы или разного масштаба. Для идентификации выбросов вычисляется расстояние Махаланобиса между каждой точкой наблюдения и центром данных. Вычисленные расстояния Махаланобиса сравниваются с распределением хи-квадрат со степенями свободы, равными числу признаков. Выбросами считались измерения больше порогового значения, определённого из распределения хи-квадрат с уровнем значимости 0,1.

Плотностной метод, основанный на кластеризации

Определение выбросов с помощью плотности также считается весьма распространённым методом. Основной принцип подхода заключается в том, что нормальные точки расположены в более плотной области точек наблюдений, в то время как выбросы находятся в более разреженных областях или вообще изолированы от остальных точек. Для задач альтиметрии был протестирован алгоритм DBSCAN. Данный алгоритм группирует тесно расположенные точки вместе и относит к выбросам точки, находящиеся в областях с малой плотностью. Преимуществом алгоритма представляется быстрая реализация для больших выборок и возможность выделения кластеров произвольной формы. Алгоритм требует задания двух гиперпараметров: ε-окрестности, определяющей область на расстоянии ε от рассматриваемой точки, и минимальное число точек, образующих плотную область (min_samples). Алгоритм начинает поиск выбросов с произвольной точки, рассчитывает её ε-окрестность, и если рассчитанная окрестность содержит достаточное число точек (не меньше min_samples), то точка и её окрестность образуют часть кластера. Все точки, находящиеся в окрестности, тоже относятся к данному кластеру. Заметим, что точка может быть позже найдена в ε-окрестности другой точки и быть включена в другой кластер. Процесс продолжается до тех пор, пока не будет найден замкнутый кластер. После этого рассматривается следующая точка из выборки. Точки, не попавшие ни в какой кластер, относятся к выбросам.

Алгоритм машинного обучения

Алгоритм «изолирующий лес», разработанный сравнительно недавно (Liu et al., 2008), использует в качестве основы модель бинарного решающего дерева. Алгоритм построения решающего дерева выглядит следующим образом:

- в корневой вершине случайным образом выбирается признак, затем случайным образом выбирается граница признака;
- часть наблюдений, у которых данный признак меньше границы, относится к левому потомку дерева, та часть, которая больше, — к правому;
- процесс рекурсивно повторяется до тех пор, пока все наблюдения не будут отделены друг от друга, либо до достижения определённых условий на дереве (например, максимальная глубина дерева).

Алгоритм требует задания двух основных гиперпараметров: 1) количество деревьев в случайном лесу (n_estimators), 2) количество признаков, необходимых для обучения каждого дерева (max_features). Дополнительно вводятся пороговый параметр (contamination) — доля выбросов в выборке и количество наблюдений для обучения каждого изолирующего дерева (max_samples). В данной работе доля выбросов была принята равной 0,1, а параметр max_samples — равным 256, как предлагается в описании алгоритма. Идея алгоритма IF заключается в том, что выбросы гораздо проще отделить от нормальных данных, поэтому расстояние от корневого узла до выброса будет существенно короче, чем до обычных данных. В случае «изолирующего леса» рассматривается усреднённое расстояние от корневого узла до наблюдения. Согласно заданной доле выбросов выбирается пороговое расстояние. Выбросом считается точка, у которой осреднённое расстояние от корневого узла будет меньше порогового значения.

Метод расчёта итоговых уровней воды

Чаще всего для практических целей спутниковой речной гидрологии используются временные серии уровней воды на виртуальных станциях (ВС). Построение временных серий состоит из двух важнейших этапов: 1) географической выборки измерений радара и соответствующей ему геофизической коррекции над руслом реки и вычислении высотной отметки уровней на момент пролёта спутника (цикл), 2) определении выбросов и осреднении в случае широких русел рек.

Получение альтиметрических измерений и геофизической коррекции из спутникового продукта

Для выборки данных из геофизического продукта спутников Sentinel-3A/B использовался снимок Landsat-8 за 17 августа 2019 г., на который накладывались все измерения радара и по которому были определены контуры русла реки в меженный период. Для получения высотных отметок уровня воды к выбранным над руслом измерениям радара были добавлены указанные выше (см. разд. «Исходные данные») геофизические коррекции.

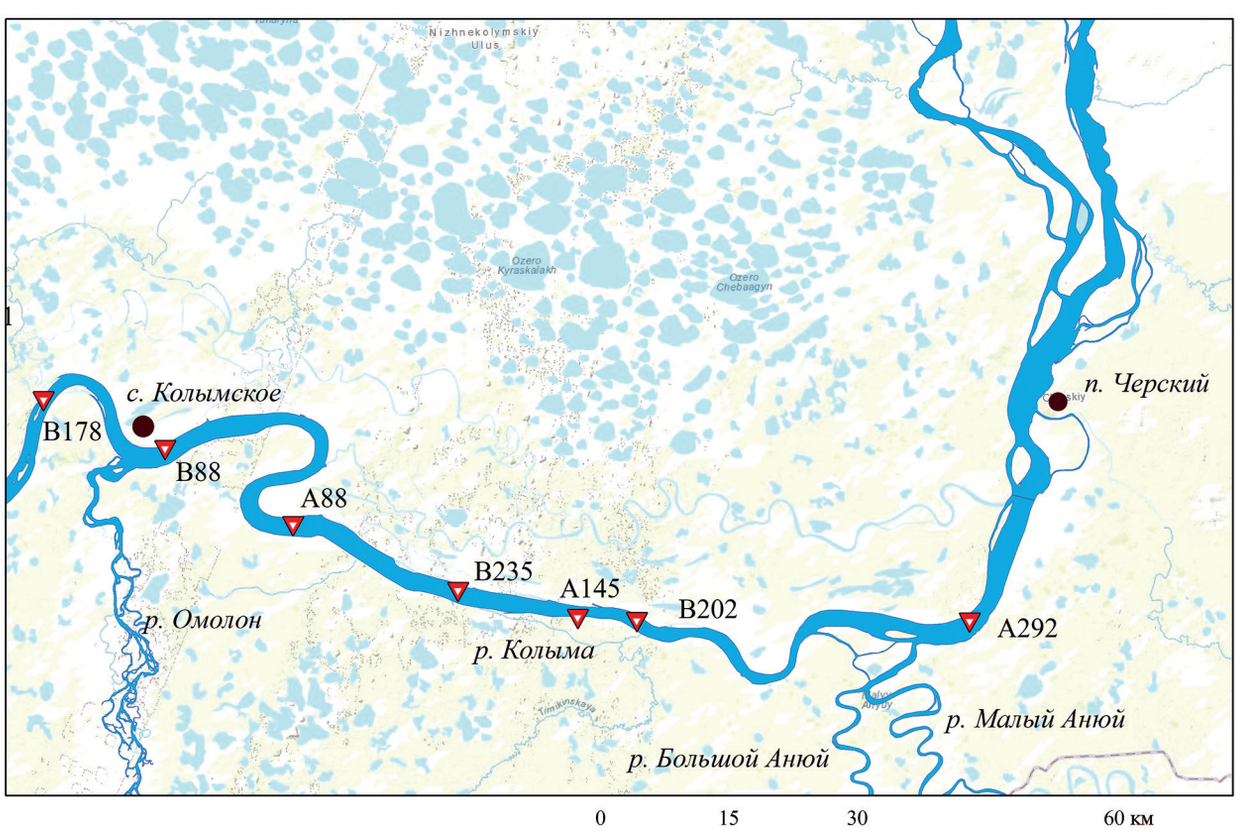


Рис. 1. Нижнее течение р. Колымы, расположение виртуальных станций и гидрометрического поста с. Колымское

Для задач исследования было отобрано семь виртуальных станций в нижнем течении р. Колымы (три виртуальные станции S3A и четыре станции S3B, *рис. 1*), которые ранее использовались для валидации гидродинамической модели, построенной на этот участок в условиях ограниченного набора натурных данных (Захарова и др., 2023). Ширина русла на участке исследования составляет 0,5–1,5 км, что подразумевает наличие 2–5 измерений уров-

ня над руслом за каждый пролёт спутника (цикл). Для каждого цикла вычислялось медианное значение уровня, принимаемое за среднее для данной даты. Полученные таким образом временные серии уровня для семи ВС получили название «исходные».

Фильтрация данных и валидация

Алгоритм фильтрации выбросов применялся к индивидуальным измерениям альтиметрического уровня (включающим 2–5 значений за цикл). В зависимости от морфологии виртуальной станции (высоты берегов, наличия стариц и осерёдков) и характера поверхности воды в русле в каждый конкретный цикл (например, ветровой ряби или волнения) в геовыборке могут оказаться как значения уровня, близкие друг к другу, так и значения, отличающиеся друг от друга на несколько метров. Случается, что только одно из пяти спутниковых измерений за данный цикл соответствует ожидаемым для данной фазы водного режима значениям уровня. Предсказать пространственное положение этого «верного» значения невозможно. Оно может наблюдаться как в середине реки, так и у берегов. Поэтому математические методы фильтрации и считаются наиболее универсальными. Предложенный в данной работе подход основан на комбинировании нескольких математических алгоритмов и состоит из следующих этапов (*рис. 2*):

1) фильтрация индивидуальных измерений с помощью правила 3σ;
2) параллельный запуск алгоритмов расстояния Махаланобиса, Isolation Forest, DBSCAN;
3) выбраковка индивидуальных измерений, которые двумя и более алгоритмами были определены как выбросы;
4) вычисление медианного значения за каждый цикл из оставшихся измерений;
5) коррекция альтиметрических серий на систематическое отклонение от отметок уровня на посту из-за разницы в моделях эллипсоида на постах и в спутниковом геофизическом продукте.

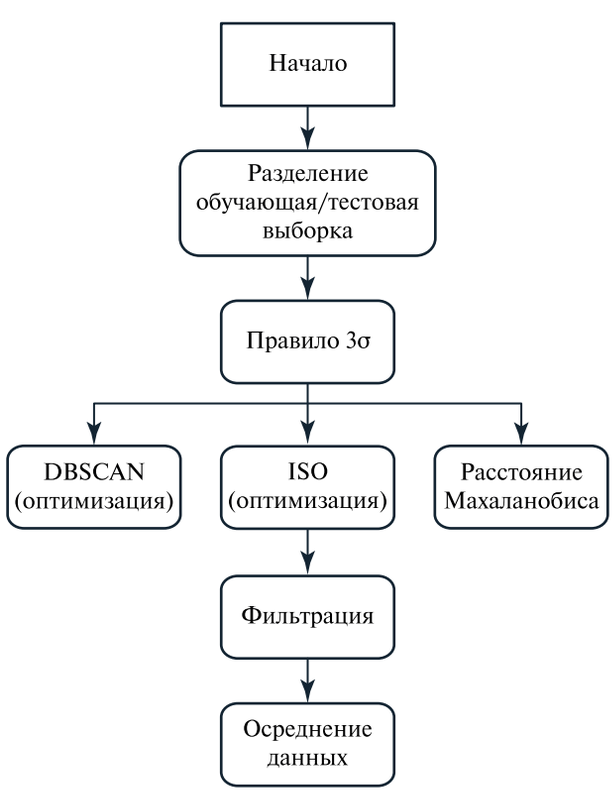


Рис. 2. Блок-схема комбинированного алгоритма построения временных серий альтиметрических уровней воды для р. Колымы

Полученные в результате временные серии уровней воды получили название «итоговые».

В алгоритмах DBSCAN и IF задаются несколько гиперпараметров, которые могут влиять на количество выбросов. Калибровка гиперпараметров проводилась с использованием натурных и спутниковых измерений для периода август 2018 г. – июль 2020 г., общего для двух спутников. Для валидации использовались данные за период август 2020 г. – сентябрь 2021 г. Эффективность фильтрации оценивалась по критерию Нэша – Сэтклиффа (*англ.* Nash-Sutcliffe model Efficiency coefficient — NSE) и среднеквадратичному отклонению (*англ.* Root Mean Square Error — RMSE) медианных значений альтиметрических уровней от уровней воды гидрометрического поста за соответствующую дату. Ввиду короткого периода наблюдений статистики рассчитывались по ансамблевой выборке, полученной из измерений на семи виртуальных станциях, соответственно за калибровочный

и валидационный периоды. Для алгоритма DBSCAN из-за известной чувствительности алгоритма к гиперпараметрам (Schubert et al., 2017) был введён дополнительный критерий, учитывающий потенциальное количество выбросов и связанный с зашумлённостью выборки спутниковых данных.

В процессе калибровки был проведён тест чувствительности алгоритмов DBSCAN и IF к изменению гиперпараметров. В качестве целевой функции выбрано среднеквадратичное отклонение.

Результаты

Временные серии уровней воды

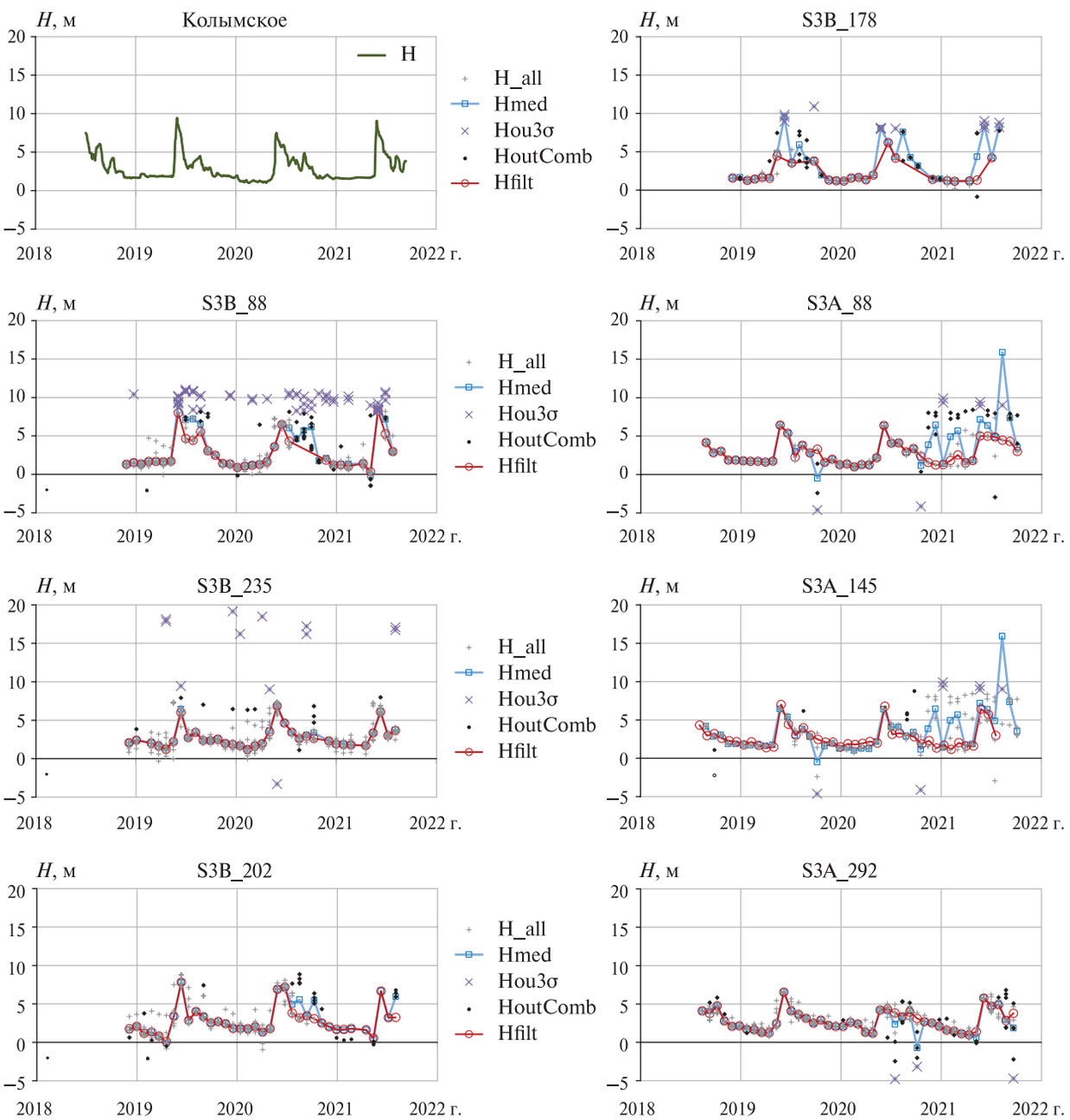


Рис. 3. Временные серии уровней воды на посту Колымское и на виртуальных станциях. Для виртуальных станций представлены исходные измерения над руслом реки (H_all), медианное значение уровня для каждого пролёта спутника (Hmed); выбросы, определённые методом 3σ (Hou3σ); выбросы, определённые комбинированным методом (HoutComb) и итоговые временные спутниковые серии уровня (Hfilt)

Как видно на *рис. 3*, несмотря на предварительный отбор спутниковых измерений по маске водной поверхности, построенной по данным спутника Landsat-8, полученные выборки уровней воды по данным альтиметрического спутника содержат как измерения, близкие к высотным отметкам водной поверхности, так и измерения, близкие к отметкам поймы. Ярким примером этому выступают виртуальные станции S3B_88, S3A_88 и S3A_145. Вдоль исследуемого участка реки превышение поймы и прирусловых валов над уровнем воды в межень колеблется в пределах 5–9 м. На рисунках, соответствующих указанным выше станциям, видно, что часть точек сгруппирована как раз около этих высот. В самом нижнем течении в пределах исследуемого участка (станции S3B_202 и S3A_292) относительные высоты правобережной поймы понижаются до 3–5 м. В результате явного кластера точек, соответствующих высотам поймы, не выявляется. Хотя на виртуальной станции S3B_202 в зимний период можно отметить группу измерений, близких к высотным отметкам правобережной поймы (~4 м). Как видно из представленных на *рис. 3* разнообразных примеров, медианное значение и статистический критерий 3σ становятся достаточно мощными инструментами для наиболее точного вычисления уровня воды рек и фильтрации очевидных выбросов. Тем не менее в ряде случаев (станции S3A_88 и S3A_145) начиная с конца 2020 г. увеличилась доля измерений, соответствующих отметкам поймы. Это привело к значительному завышению итоговых медианных уровней воды. Только наиболее высокие значения «пойменных» уровней были отфильтрованы с помощью статистического метода 3σ. Алгоритмы DBSCAN и «изолирующий лес» позволили эффективно удалить оставшиеся точки, не соответствующие уровню реки, на станции S3A_88. Но оба алгоритма оказались неэффективны на станции S3A_145. Несмотря на их неэффективность, оставшиеся единичные «пойменные» значения не повлияли на вычисления итогового медианного уровня воды на данной станции, так как наиболее значимые выбросы были удалены с помощью критерия 3σ. В *табл. 1* представлены результаты сравнения ансамблевых выборок исходных и итоговых значений уровней воды по данным альтиметра (совмещённые измерения на семи виртуальных станциях) с натурными наблюдениями за уровнями воды на водомерном посту Колымское для периодов калибровки и валидации.

Таблица 1. Оценка точности рядов альтиметрических наблюдений (RMSE и критерий Нэша – Сэтклиффа для ансамблевых выборок, полученных при осреднении исходных измерений, при осреднении после фильтрации методом 3σ и после фильтрации комбинированным методом)

	Исходные серии		Фильтр 3σ		Комбинированный фильтр	
	RMSE, м	NSE	RMSE, м	NSE	RMSE, м	NSE
Период калибровки гиперпараметров (2019–2020)	0,91	0,72	0,84	0,76	0,64	0,82
Период валидации (август 2020 г. – сентябрь 2021 г.)	1,77	–0,24	1,42	0,20	0,67	0,80
Расширенный период валидации для станций S3A (2016–2018 гг., август 2019 г. – сентябрь 2021 г.)	1,81	–0,33	1,17	0,44	0,69	0,68

Для виртуальных станций спутника Sentinel-3A период валидации может быть продлён за счёт включения дополнительных 2,5 лет наблюдений до запуска Sentinel-3B (2016–2018). Несмотря на некоторое ухудшение точности итоговой ансамблевой выборки для расширенного периода наблюдений на станциях Sentinel-3A, в целом комбинированный метод продемонстрировал достаточно высокую эффективность в отношении определения выбросов и за указанный предшествующий период. Эффективность фильтрации выбросов с использованием простого статистического и более сложного комбинированного алгоритма, включающего метод машинного обучения, для каждой виртуальной станции представлена в *табл. 2.*

Таблица 2. Оценка точности рядов альтиметрических измерений до и после фильтрации на индивидуальных виртуальных станциях для периода валидации (август 2020 г. – сентябрь 2021 г.)

Трек	Исходные данные				3σ				Комбинированный алгоритм			
	N_{obs}	RMSE, м	NSE	bias	N_{obs}	RMSE, м	NSE	bias	N_{obs}	RMSE, м	NSE	bias
A88	68	2,99	–3,85	4,82	60	1,64	–0,46	6,96	38	0,72	0,72	6,26
A145	40	1,19	0,03	6,75	39	1,09	0,59	7,53	34	0,71	0,83	6,90
A292	121	1,88	0,37	7,69	117	0,89	0,66	5,64	99	0,85	0,69	7,55
B88	133	1,20	0,56	6,11	104	0,69	0,85	6,58	85	0,49	0,84	6,04
B178	39	1,88	–0,32	5,3	34	1,87	–1,82	5,38	21	0,28	0,93	5,68
B202	93	1,41	0,17	6,93	93	1,41	0,17	6,93	74	0,58	0,86	7,16
B235	137	0,75	0,76	6,85	133	0,76	0,76	6,87	127	0,70	0,80	6,89

Примечание: bias — систематическое отклонение; N_{obs} — количество измерений на каждой виртуальной станции.

При использовании комбинированного алгоритма общее количество выбросов из ансамблевой выборки за период калибровки составило 13 %, а за период валидации —24 %. Количество выбросов для индивидуальных виртуальных станций для валидационного периода варьировало от 7 до 46 %. При этом вклад комбинированного подхода в количество определённых выбросов составил от 4 до 33%.

Тестирование чувствительности алгоритмов фильтрации выбросов к настройке гиперпараметров

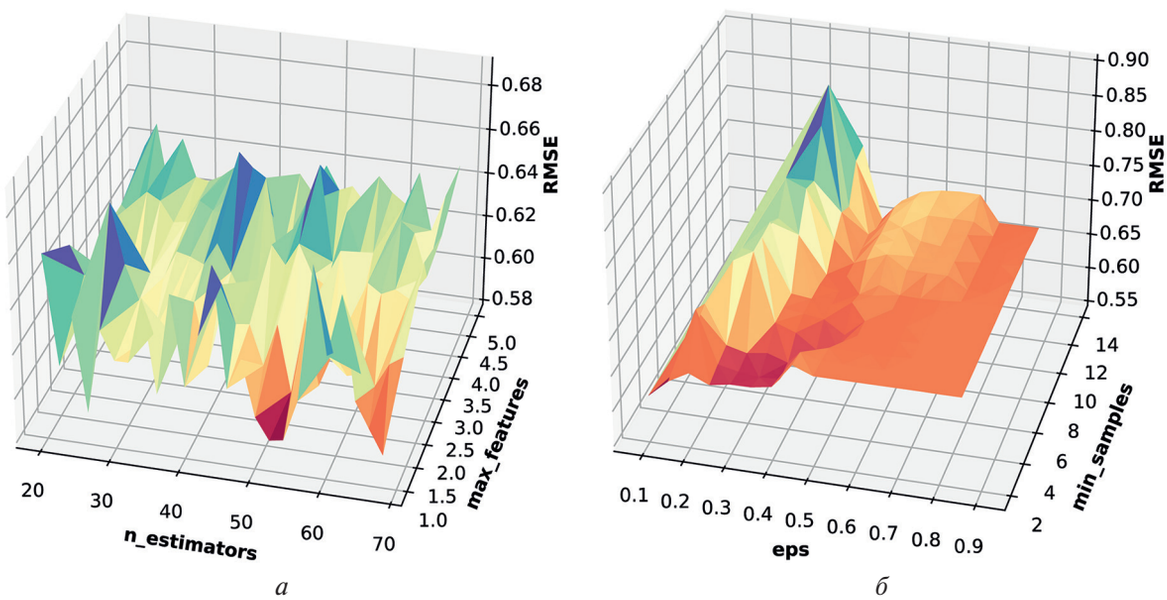
Для ансамблевой выборки спутниковых измерений комбинированный алгоритм фильтрации спутниковых альтиметрических измерений позволил увеличить точность расчётов уровней воды в нижнем течении р. Колымы на 0,48–0,75 м по сравнению с простым статистическим методом фильтрации (3σ). Не исключено, что комбинированный алгоритм позволил бы добиться более высоких результатов, однако при этом, возможно, пришлось бы потерять спутниковые наблюдения за отдельные даты, как произошло, например, на станции S3B_88, где измерения во время паводка 2020 г. были полностью забракованы. Тест влияния выбора гиперпараметров алгоритмов DBSCAN и IF на результаты фильтрации продемонстрировал, что алгоритм IF мало чувствителен к выбору гиперпараметров (*рис. 4а*, см. с. 85). При увеличении значений гиперпараметров в 2–3 раза по сравнению с найденными оптимальными значениями (*табл. 3*) значение RMSE в выбранном диапазоне тестирования варьировалось от 0,59 до 0,67 м.

Таблица 3. Набор оптимальных гиперпарметров, полученных при калибровке на ансамблевой выборке

Алгоритм	Параметр 1	Параметр 2	Значение RMSE
DBSCAN	ε = 0,65	min_samples = 4	0,67
IF	n_estimators = 52	max_features = 2	0,60

Алгоритм DBSCAN известен своей чувствительностью к выбору гиперпараметров. В процессе оптимизации гиперпараметров DBSCAN и тестирования алгоритма на чувствительность к выбору гиперпараметров было выявлено, что оптимальное решение, удовлетворяющее условию «минимум RMSE», лежит в зоне низких значений ε и min_samples и высокой вариабельности объективной функции RMSE (*рис. 4б*). В то же время в зоне средних значений указанных гиперпараметров вариабельность RMSE незначительна. В целях более стабильной

работы алгоритма и во избежание переобучения было решено в качестве оптимального набора гиперпараметров DBSCAN выбрать ε и min_samples, лежащие в районе плато, со значениями ε более 0,5 и min_samples менее 10 точек (*табл. 3*). Значения RMSE на данном плато немного выше и варьируют от 0,67 до 0,68 м. При этом подходе уменьшение точности итоговой временной серии ансамблевой выборки для периода калибровки составило 0,07 м по сравнению с наборами оптимальных параметров, лежащих в зоне минимальных значений RMSE.



а *б*

Рис. 4. Чувствительность моделей фильтрации выбросов IF (*а*) и DBSCAN (*б*) в зависимости от гиперпараметров

Заключение и перспективы

Предложенный комбинированный метод фильтрации индивидуальных значений альтиметрических измерений спутников Sentinel-3A и Sentinel-3B позволил значительно увеличить точность расчётов итоговых временных серий уровня воды для сложных морфологических условий широкопойменного участка арктической реки. Для ансамблевой выборки средних за цикл значений уровня воды улучшение точности составило 40–50 % по сравнению с выборкой, полученной при использовании стандартного правила фильтрации выбросов 3σ. Для индивидуальных виртуальных станций ошибка в определённых по данным альтиметрии уровнях воды уменьшилась на 4–85 %, или 0,04–1,59 м. Мы считаем разработанный метод фильтрации данных достаточно перспективным. Решение о выбросе принимается при условии, когда два из трёх алгоритмов (расстояния Махаланобиса, DBSCAN и «изолирующий лес») классифицируют рассматриваемое измерение как выброс. Данный подход позволяет уменьшить влияние потенциальных индивидуальных недостатков каждого из трёх алгоритмов. Несмотря на то что алгоритм оказался достаточно строгим к входящим данным (например, он исключил ряд значений в период паводка в 2020 г. для станций S3B_88, S3B_178), большинство данных было отфильтровано корректно. Стоит отметить, что оставшиеся финальные значения являются значениями, отражающими ход фактического уровня воды с высокой степенью достоверности. В дальнейшем, при увеличении количества лет наблюдений, будет возможно разделение данных по сезонной амплитуде на две группы и применение алгоритма к каждой группе отдельно, что должно привести к улучшению результатов. В перспективе разработанный подход будет протестирован на других широкопойменных участках арктических рек, для которых в настоящее время точность спутниковых измерений уровней воды ниже, чем для узкопойменных рек или рек умеренного и тропического поясов. Изменчивость гиперпараметров алгоритмов DBSCAN и «изолирующий лес» в зависимости

от морфологических условий участков рек (узкопойменные однорукавные, многорукавные, при наличии высоких пойменных террас и т. д.) или от спутникового инструмента планируется изучить в ближайший год. С учётом последнего интересной задачей представляется сравнительный анализ эффективности работы разработанного метода с измерениями Sentinel-3A/B и другого альтиметрического спутника — Jason-3, периодичность измерений которого (10 дней) более адаптирована для решения гидрологических задач.

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 22-27-00633 «Исследование уровенного режима рек методами спутниковой альтиметрии и гидродинамического моделирования»).

Литература

1. *Захарова Е. А.*, *Крыленко И. Н.*, *Сазонов А. А.*, *Семенова Н. К.*, *Лисина А. А.* Уровенный режим арктических рек по данным моделирования и спутниковых измерений // Метеорология и гидрология. 2023. № 12. С. 115–124.
2. *Abdalla S.*, *Kolahchi A. A.*, *Ablain M. et al.* Altimetry for the future: Building on 25 years of progress // Advances in Space Research. 2021. V. 68. P. 319–363. DOI: 10.1201/9781315151779-5.
3. ATBD: Algorithm Theoretical Basis Document, Deliverable D1.3, Sentinel-3 and Cryosat SAR/SARin Radar Altimetry for Coastal Zone and Inland Water. ESA Contract. 2022. 4000129872/20/I-DT. 123 p.
4. *Biancamaria S.*, *Schaedele T.*, *Blumstein D. et al.* Validation of Jason-3 tracking modes over French rivers // Remote Sensing of Environment. 2018. V. 209. P. 77–89. DOI: 10.1016/j.rse.2018.02.037.
5. *Liu F. T.*, *Ting K. M.*, *Zhou Z. H.* Isolation Forest // 8th IEEE Intern. Conf. Data Mining (ICDM'08). 2008. P. 413–422. DOI: 10.1109/ICDM.2008.17.
6. *Maillard P.*, *Bercher N.*, *Calmant S.* New processing approaches on the retrieval of water levels in Envisat and SARAL radar altimetry over rivers: A case study of the Sao Francisco River, Brazil // Remote Sensing of Environment. 2015. V. 156. P. 226–241. DOI: 10.1016/j.rse.2014.09.027.
7. *Rémy F.*, *Flament T.*, *Blarel F.*, *Benveniste J.* Radar altimetry measurements over Antarctic ice sheet: a focus on antenna polarization and change in backscatter problems // Advances in Space Research. 2012. V. 50. P. 998–1006. DOI: 10.1016/j.asr.2012.04.003.
8. *Schubert E.*, *Sander J.*, *Ester M.*, *Kriegel H.-P.*, *Xu X.* DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN // ACM Trans. Database Systems. 2017. V. 42. P. 1–21. DOI: 10.1145/3068335.
9. *Schwatke C.*, *Dettmering D.*, *Bosch W.*, *Seitz F.* DAHITI — an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry // Hydrology and Earth System Sciences. 2015. V. 19. P. 4345–4364. DOI: 10.5194/hess-19-4345-2015.
10. *Zakharova E.*, *Nielsen K.*, *Kamenev G.*, *Kouraev A.* River discharge estimation from radar altimetry: Assessment of satellite performance, river scales and methods // J. Hydrology. 2020. V. 583. Article 124561. DOI: 10.1016/j.jhydrol.2020.124561.

Statistical and machine learning methods for river water level time series construction using satellite altimetry

N. K. Semenova [1,2], E. A. Zakharova [1,3], I. N. Krylenko [1,2], A. A. Sazonov [1,2]

[1] *Water Problem Institute RAS, Moscow 119333, Russia*
E-mail: snkone132@mail.ru
[2] *Lomonosov Moscow State University, Moscow 119991, Russia*
[3] *Earth Observation for Learning and Application, Toulouse 31400, France*

The use of satellite altimetry data for monitoring the level regime of rivers in Arctic regions is limited due to the negative effect of complex fluvial morphology and ice cover on altimetric radar mea-

surements. The construction of time series of river water levels consists of two main stages: 1) accurate geographic sampling of satellite measurements over the river channel and 2) calculation of the average level for a given date after filtering outliers. This work is based on measurements from the European altimetry satellites Sentintel-3A and Sentinel-3B. The paper proposes a method for determining aberrant values of altimetric measurements (outliers) over the wide floodplain section of the Kolyma River. The method improved the accuracy of calculation of satellite time series of water level by 0.04–1.59 m (or 4–85 %) compared to the widely used standard statistical method of filtering altimetric measurements. The suggested method is based on the combination of three algorithms of different complexity: statistical (Mahalanobis distance), clustering (Density-Based Spatial Clustering of Applications with Noise (DBSCAN)) and machine learning (Isolating Forest) methods. In the combined approach, values classified as outliers by at least two algorithms were considered outliers. This approach allowed us to reduce the impact of potential individual shortcomings of each of the three methods.

Keywords: satellite altimetry, Arctic rivers, water level, detection of outliers, machine learning methods

Accepted: 08.12.2023
DOI: 10.21046/2070-7401-2024-21-1-76-87

References

1. Zakharova E. A., Krylenko I. N., Sazonov A. A., Semenova N. K., Lisina A. A., Water level regime of Arctic rivers according to modeling and satellite measurements, *Meteorologiya i gidrologiya*, 2023, No. 12, pp. 115–124 (in Russian).
2. Abdalla S., Kolahchi A. A., Ablain M. et al., Altimetry for the future: Building on 25 years of progress, *Advances in Space Research*, 2021, Vol. 68, pp. 319–363, DOI: 10.1201/9781315151779-5.
3. *ATBD: Algorithm Theoretical Basis Document*, *Deliverable D1.3*, *Sentinel-3 and Cryosat SAR/SARin Radar Altimetry for Coastal Zone and Inland Water*, ESA Contract, 2022, 4000129872/20/I-DT, 123 p.
4. Biancamaria S., Schaedele T., Blumstein D. et al., Validation of Jason-3 tracking modes over French rivers, *Remote Sensing of Environment*, 2018, Vol. 209, pp. 77–89, DOI: 10.1016/j.rse.2018.02.037.
5. Liu F. T., Ting K. M., Zhou Z. H., Isolation Forest, *8th IEEE Intern. Conf. Data Mining* (*ICDM'08*), 2008, pp. 413–422, DOI: 10.1109/ICDM.2008.17.
6. Maillard P., Bercher N., Calmant S., New processing approaches on the retrieval of water levels in Envisat and SARAL radar altimetry over rivers: A case study of the Sao Francisco River, Brazil, *Remote Sensing of Environment*, 2015, Vol. 156, pp. 226–241, DOI: 10.1016/j.rse.2014.09.027.
7. Rémy F., Flament T., Blarel F., Benveniste J., Radar altimetry measurements over Antarctic ice sheet: a focus on antenna polarization and change in backscatter problems, *Advances in Space Research*, 2012, Vol. 50, pp. 998–1006, DOI: 10.1016/j.asr.2012.04.003.
8. Schubert E., Sander J., Ester M., Kriegel H.-P., Xu X., DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, *ACM Trans. Database Systems*, 2017, Vol. 42, pp. 1–21, DOI: 10.1145/3068335.
9. Schwatke C., Dettmering D., Bosch W., Seitz F., DAHITI — an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry, *Hydrology and Earth System Sciences*, 2015, Vol. 19, pp. 4345–4364, DOI: 10.5194/hess-19-4345-2015.
10. Zakharova E., Nielsen K., Kamenev G., Kouraev A., River discharge estimation from radar altimetry: Assessment of satellite performance, river scales and methods, *J. Hydrology*, 2020, Vol. 583, Article 124561, DOI: 10.1016/j.jhydrol.2020.124561.