

Monthly Runoff Forecasting by Non-Generalizing Machine Learning Model and Feature Space Transformation (Vakhsh River Case Study)

Matrenin P.V.¹, Safaraliev M.K.², Kiryanova N.G.¹, Sulonov S.M.³

¹Novosibirsk State Technical University, Novosibirsk, Russian Federation

²Ural Federal University, Yekaterinburg, Russian Federation

³Tajik Technical University, Dushanbe, Tajikistan

Abstract. Energy prices and cost of materials for solar and wind power plants have increased over the past year. Therefore, significance increases for the hydropower and long-term (1–10 years) planning generation for the existing hydropower plants, which requires forecasting the average monthly values of the river flow. This task is especially urgent for countries without their own oil-fields and opportunity to invest in the creation of solar or wind power plants. The aim of the research is to decrease the mean absolute forecasting error of the long-term prediction for the Vakhsh River flow (Tajikistan) based on the long-term observations. A study of existing methods for the river runoff forecasting in relation to the object under consideration was carried out, and a new transformation model for the space of the input features was developed. The most significant results are the decrease in the average forecast error in the Vakhsh river flow achieved by the use of the proposed space of polynomial logarithmic features in comparison with other methods, and the need to use at least the 20 year-old observational data for the long-term operation planning of the hydropower plants and cascades of the hydropower plants obtained from the results of computational experiments. The significance of the results lies in the fact that a new approach to the long-term forecasting of river flow has been proposed and verified using the long-term observations. This approach does not require the use of the long-term meteorological forecasts, which are not possible to obtain with high accuracy for all regions.

Keywords: river flow, hydropower, long-term forecasting, generation planning, machine learning, Republic of Tajikistan, cascade of hydropower plants.

UDC: 621.311: 519.25

DOI: <https://doi.org/10.52254/1857-0070.2022.3-55.08>

Predicția valorilor medii lunare ale debitelor râurilor folosind un model de învățare automată negeneralizat și transformarea spațiului de caracteristici (pe exemplul râului Vahş)

Matrenin P.V.¹, Safaraliev M.H.², Kiryanova N.G.¹, Sulonov Ş.M.³

¹Universitatea Națională tehnică din Novosibirsk, Novosibirsk, Federația Rusă

²Universitatea Națională Federală, Ecaterinburg, Federația Rusă

³Universitatea Tehnică din Tadjikistan, Duşanbă, Tadjikistan

Rezumat. În ultimul an, prețurile atât pentru purtătorii de energie cu hidrocarburi, cât și pentru materialele utilizate pentru producția de centrale solare și eoliene au crescut vertiginos, în plus, a existat o lipsă de energie electrică în sistemele energetice, cu o pondere mare a centralelor solare și eoliene din cauza condițiilor meteorologice anormale din anumite regiuni. Importanța hidroenergiei și a planificării pe termen lung (1–10 ani) pentru generarea hidrocentralelor existente este în creștere, ceea ce, la rândul său, necesită prognozarea valorilor medii lunare ale scurgerii râului. Scopul lucrării este de a reduce eroarea medie de prognoză pe termen lung a scurgerii râului Vakhsh (Republica Tadjikistan) pe baza observațiilor pe termen lung. Pentru atingerea scopului, a fost realizat un studiu al metodelor existente de predicție a debitelor râurilor în raport cu obiectul luat în considerare și a fost dezvoltat un nou model de transformare a spațiului caracteristicilor de intrare. Cele mai importante rezultatele sunt scăderea erorii medii de prognoză a scurgerii râului Vakhsh în comparație cu alte metode, realizată prin utilizarea spațiului de caracteristici logaritmice polinomiale și justificarea obținută din rezultatele experimentelor de calcul pentru necesitatea utilizării datelor pentru cel puțin 20 de ani pentru planificarea pe termen lung a exploatarea hidrocentralelor și cascadelor hidrocentralelor. Semnificația rezultatelor constă în faptul că a fost propusă și verificată o nouă abordare a prognozării pe termen lung a debitului râurilor pe baza urmării datele pe termen lung, care nu necesită utilizarea prognozelor meteorologice pe termen lung, care nu sunt posibile de obținut cu exactitate pentru toate regiunile.

Cuvinte-cheie: debit fluvial, energie hidroelectrică, prognoză pe termen lung, planificare a producției, învățare automată, Republica Tadjikistan, cascadă hidroelectrică.

**Прогнозирование среднемесячных значений стоков рек с применением необобщающей модели машинного обучения и преобразованием пространства признаков (на примере реки Вахш)
Матренин П.В.¹, Сафаралиев М.Х.², Кирьянова Н.Г.¹, Султонов Ш.М.³**

¹Новосибирский государственный технический университет, Новосибирск, Российская Федерация

²Уральский федеральный университет, Екатеринбург, Российская Федерация

³Таджикский технический университет, Душанбе, Таджикистан

Аннотация. За последний год стремительно выросли цены как на углеводородные энергоносители, так и на материалы, используемые для производства солнечных и ветровых электростанций, кроме того, наблюдался дефицит электроэнергии в энергетических системах с высокой долей солнечных и ветровых электрических станций из-за аномальных погодных условий в отдельных регионах. Поэтому увеличивается значимость гидроэнергетики и долгосрочного (1–10 лет) планирования выработки существующих гидроэлектростанций, что, в свою очередь, требует прогнозирования среднемесячных значений речного стока. Особенно эта задача актуальна для стран, не имеющих собственных нефтяных месторождений и возможности инвестирования в создание солнечных или ветровых электростанций. Целью работы является снижение средней ошибки долгосрочного прогнозирования стока реки Вахш (Республика Таджикистан) по данным многолетних наблюдений. Для достижения цели было проведено исследование существующих методов прогнозирования стоков реки применительно к рассматриваемому объекту и разработанная новая модель преобразования пространства входных признаков. Наиболее существенными результатами являются достигнутое за счет использования пространства полиномиальных логарифмированных признаков снижение средней ошибки прогнозирования стока реки Вахш по сравнению с другими методами и полученное по результатам вычислительных экспериментов обоснование необходимости использовать данные наблюдений как минимум за 20 лет для долгосрочного планирования работы гидроэлектростанций и каскадов гидроэлектростанций. В настоящее время на реке Вахш расположено восемь гидроэлектростанций, которые составляют 85 % всей установленной мощности энергосистемы Таджикистана, поэтому точность прогнозирования стока оказывает большое влияние на планирование работы всей энергосистемы страны. Значимость результатов заключается в том, что предложен и верифицирован на данных многолетних наблюдений новый подход к долгосрочному прогнозированию речного стока, который не требует использования долгосрочных метеорологических прогнозов, получение которых с высокой точностью возможно не для всех регионов.

Ключевые слова: сток реки, гидроэнергетика, долгосрочное прогнозирование, планирование выработки, машинное обучение, Республика Таджикистан, каскад гидроэлектростанций.

ВВЕДЕНИЕ

А. Задача прогнозирования речного стока

В общей сложности на гидроэнергетику приходится более 17 % мирового производства электроэнергии, которая на сегодняшний день является одним из наиболее широко используемых экологически чистых источников энергии [1]. Это в первую очередь связано с более низкой стоимостью производства энергии с помощью данного источника, особенно по сравнению с тепловой энергией [2].

Крупные гидроэлектростанции (ГЭС) используются для выработки электроэнергии с целью удовлетворения пиковых потребностей энергосистемы [3]. Технологии использования возобновляемых источников энергии вносят значительный вклад в сокращение выбросов парниковых газов и обеспечение безопасности энергоснабжения. Так, по сравнению с обычными угольными электростанциями, гидроэнергетика предотвращает выброс около 3 Гт CO₂ в год, что составляет около 9% мировых ежегодных выбросов CO₂ [4].

В то же время важным недостатком использования гидроэнергии является необходимость планирования режимов работы ГЭС с учетом правил использования водных ресурсов, что осложняет не только планирование тарифов, бюджетов на год и инвестиций, но и переговоры по контрактам в конкретном месяце [1]. Поэтому точное прогнозирование стока рек имеет решающее значение для эффективного управления водными ресурсами, включающего в себя помимо планирования производства гидроэлектроэнергии планирование орошения, прогнозирование наводнений и других гидрологических процессов. Из-за нелинейного поведения временных рядов потока прогнозирование стока реки остается одним из наиболее сложных вопросов в области гидрологических наук [5-7].

Исследования показывают, что наибольшую точность показывают физико-математические модели, учитывающие большое число параметров: цифровые модели рельефа, данные о почвах территории бассейна реки, многолетние наблюдения метеорологических и гидрологических параметров, данные дистанционного зондирования Земли [6-10].

Можно выделить модель ECOMAG, включающую помимо физико-математических моделей средства для геоинформационного моделирования, уникальные базы многолетних гидрометеорологических данных, базы данных о пространственном распределении видов почв, растительности, климате и другие [6] и комплекс на базе ГИС-технологий «GRASS GIS» [9]. Недостатками данных моделей являются, во-первых, высокая сложность и трудоемкость применения, во-вторых, необходимость большого количества данных, которые доступны не для всех водных объектов [6]. Иными словами, применение подобных моделей является научной задачей, требующей высокой квалификации исполнителя и нескольких лет исследований. Кроме того, подобные комплексы не доступны для широкого использования, так как или имеют высокую цену или в принципе не распространяются за пределы института, разрабатывающего их. Необходимые данные не всегда доступны для регионов, в которых не проводилось глубоких геологических, гидрологических, метеорологических исследований или не собраны данные метеорологических и гидрологических наблюдений за 20 и более лет.

Альтернативным подходом являются модели «черного ящика», в которых не используется создание физически обоснованной модели, учитывающей рельеф, почвы, растительность. Вместо этого методами статистического анализа или машинного обучения строится функциональная зависимость между требуемым прогнозным значением и теми входными данными, которые имеются для рассматриваемого объекта [8, 11, 12].

В задачах прогнозирования стока применяется большое количество методов машинного обучения [11, 12], их выбор во многом зависит от того, какие имеются исходные данные и в каком объеме, от горизонта планирования и от особенностей самой реки. В данном исследовании рассмотрена методика создания непараметрических моделей на основе алгоритма k -ближайших соседей (k -Nearest Neighbor, kNN).

В. Непараметрическая модель kNN

Алгоритм k -ближайших соседей [13] является одним из наиболее распространенных алгоритмов машинного обучения. Его достоинства – простота реализации и прозрачная процедура работы, что приводит к высокой интер-

претируемости результатов, в отличие от многих методов машинного обучения, которые создают слабо интерпретируемые модели или не интерпретируемые модели, таких как ансамбли деревьев решений, нейронные сети, метод опорных векторов. Недостатком kNN является необходимость выбора гипер-параметров, в первую очередь числа соседей k , а также метрики расстояний (distance function). Большинство работ, посвященных kNN , фокусируются на подборе гипер-параметров, введении в алгоритм принципов нечетких множеств или на повышении скорости работы алгоритма и снижении используемой памяти за счет ускорения поиска ближайших соседей. Оптимизация по скорости работы и по памяти необходима только в задачах очень высокой размерности [14], рассматриваемая задача долгосрочного прогнозирования такой не является. К первой группе можно отнести работы, которые рассматривают методы поиска гипер-параметров на основе поиска по сетке (Grid Search) и случайного поиска (Random Search) [15]; исследования, предлагающие определять число k динамически для каждого рассматриваемого объекта [16] или использовать адаптивные функции для метрики расстояний и функции расчета весовых коэффициентов соседей [17, 18]. Кроме того, можно выделить нечеткие модификации kNN , представленные в работах [19, 20], и различные гибридные модели на базе kNN и нейронных сетей [21].

В задачах небольшой размерности выбор оптимального числа k путем перебора по сетке или случайного перебора не занимает много времени, при этом такой подход никак не усложняет сам алгоритм и не снижает простоты интерпретации результатов его работы, а значит не повышает риск переобучения или подгонки под данные (overfitting). Методы, адаптирующие гипер-параметры алгоритма под данные, усложняют его и повышают риск переобучения за счет настройки адаптации под определенную выборку данных. Гибридизация kNN с более сложными моделями значительно снижает интерпретируемость алгоритма, а при использовании нейронных сетей возникает и проблема слабой обучаемости нейронных сетей при малом объеме данных.

В данной работе используется подход, не усложняющий классический алгоритм kNN , при этом повышающий его точность в задаче регрессии за счет этапа извлечения признаков (feature-extraction). Подход аналогичен тому,

который используется для повышения точности линейной регрессии с помощью построения полиномиальных комбинаций признаков (при этом сама модель остается линейной, меняется только пространство признаков).

Цель данной работы – повышение точности долгосрочного (от 1 года) прогнозирования стока реки Вахш, важной для обеспечения электроэнергией Республики Таджикистан [22, 23], за счет применения непараметрической модели машинного обучения к данным многолетних наблюдений. Отличием работы является создание модификации алгоритма kNN на основе полиномиального логарифмического преобразования признаков и исследование влияния длительности используемых наблюдений на точность прогнозирования стока реки.

В работе также проведено сравнение результатов с другими методами машинного обучения, статистическими авторегрессионными моделями и подходом на базе метеорологических факторов. Применение физико-математических моделей для рассматриваемого объекта затруднено из-за отсутствия необходимых данных (состав почт, рельеф, растительность). Для повышения точности kNN в работе предложен новый алгоритм преобразования пространства входных признаков.

II. МЕТОДЫ ИССЛЕДОВАНИЯ

A. Особенности объекта исследования и исходные данные

Река Вахш считается одной из важнейших рек в Таджикистане, с точки зрения потенциала производства электроэнергии, и вторым по величине северо-западным притоком реки Амударья в бассейне Аральского моря в Центральной Азии и образуется от слияния рек Сурхоб и Оби-Хингоу. Река Вахш является основным правым притоком реки Амударья, длина реки составляет 524 км, площадь бассейна 39,1 тыс. км². Большая часть бассейна реки находится в горной системе Памиро-Алая. На реке Вахш каскадом расположены восемь ГЭС. Шесть из них находятся на самой реке Вахш: Строищаяся Рогунская, Нурекская, Байпазинская, Сангтудинская 1, Сангтудинская 2 и Головная ГЭС. Две из них Центральная и Перепадная находятся на магистральном Вахшском канале. Более 90 % установленной мощности энергосистемы Республики Таджикистана, приходится на ГЭС,

около 95 % процентов сосредоточена на реке Вахш [22, 23]. Бассейн расположен между 37,10 и 39,74 северной широты и 68,31 и 73,70 восточной долготы. Высота над уровнем моря в бассейне колеблется от 302 до 7050 м над уровнем моря, Рисунок 1. Для Вахша характерно низкое состояние уровней и расходов в осенне-зимний период, когда питание реки осуществляется в основном грунтовыми водами и периодически выпадающими осадками. Подъем расходов воды начинается в апреле, наибольшие расходы воды наблюдаются в июле, иногда в конце или начале августа, с середины августа начинается спад, продолжающийся до октября [24, 25].

В верховьях реки Вахш из-за ограничений в наличие подходящей земли, орошение довольно ограничено. Кроме того, вода для этой мелкомасштабной ирригационной установки берется из притоков реки Вахш, а не берется непосредственно из самой реки. Следовательно, вода, используемая для этой цели, не подтверждается измерением стока реки.

В исследовании использованы среднемесячные значения стока реки Вахш (Таджикистан) за период с января 1927 по декабрь 2015 гг. (89 лет), фрагменты выборки приведены на Рисунках 2, 3. Из них видно, что Вахш – очень сезонная горная река, с максимумом стока в июле и минимумом в феврале. Река течет в основном по узкой долине, местами переходящей в непроходимые ущелья шириной 8-10 м, а в некоторых местах расширяется до 1,5 км, на ее сток в основном влияет таяние снега, поскольку большая часть годовых осадков выпадает в зимние месяцы, в более высоких районах, в виде снега.

Сезонность стока что позволяет предположить, что для задачи эффективны будут авторегрессионные модели, такие как ARIMA (autoregressive integrated moving average) и SARIMA (seasonal ARIMA) [26, 27].

Для моделей машинного обучения очень важен выбор признаков, поэтому построен график изменения коэффициентов корреляции Пирсона между стоком в определенный месяц и предыдущие месяцы на интервале 30 лет, показанный на Рисунке 4. Очевидно, что наиболее значимыми являются данные за аналогичные месяцы предыдущих лет, а также за противоположные месяцы, например, март-сентябрь.

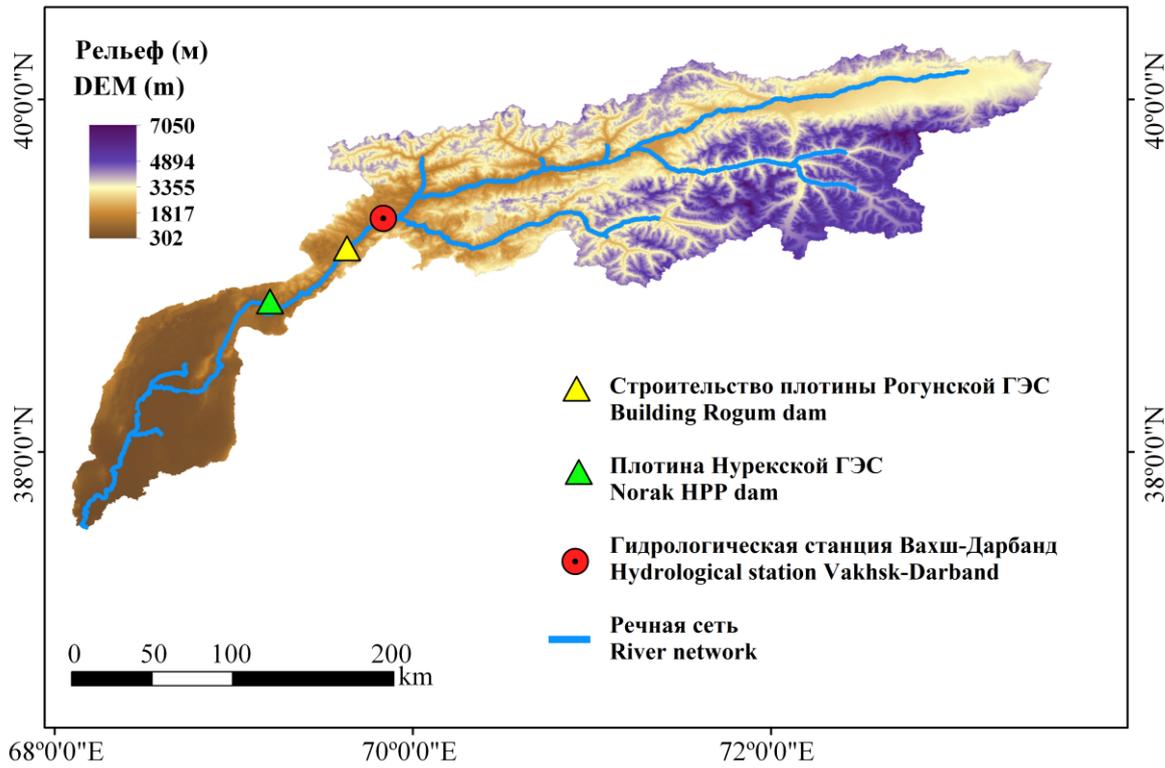


Рис.1. Географическое расположение исследуемого региона.
 Fig. 1. Geographic location of the study region.

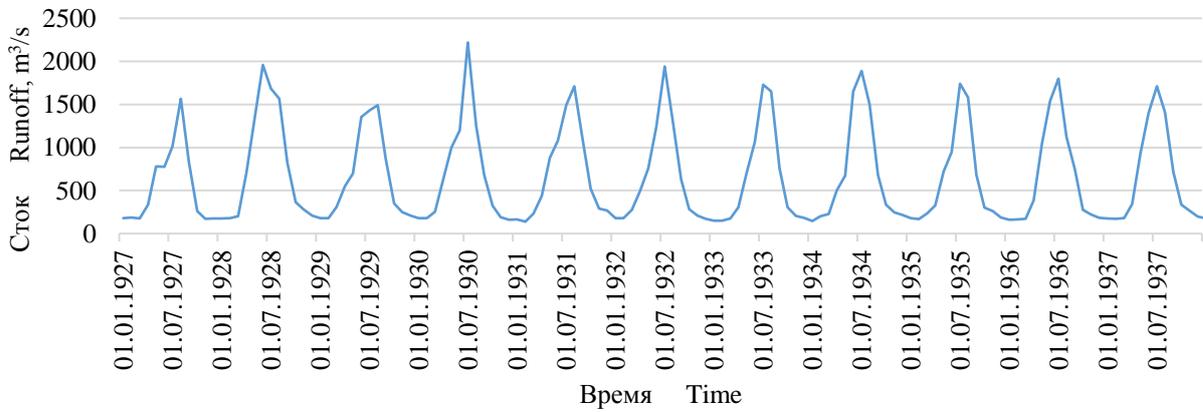


Рис.2. Сток реки Вахш 1927-1937 гг.
 Fig. 2. Vakhsh River runoff for 1927-1937 years.

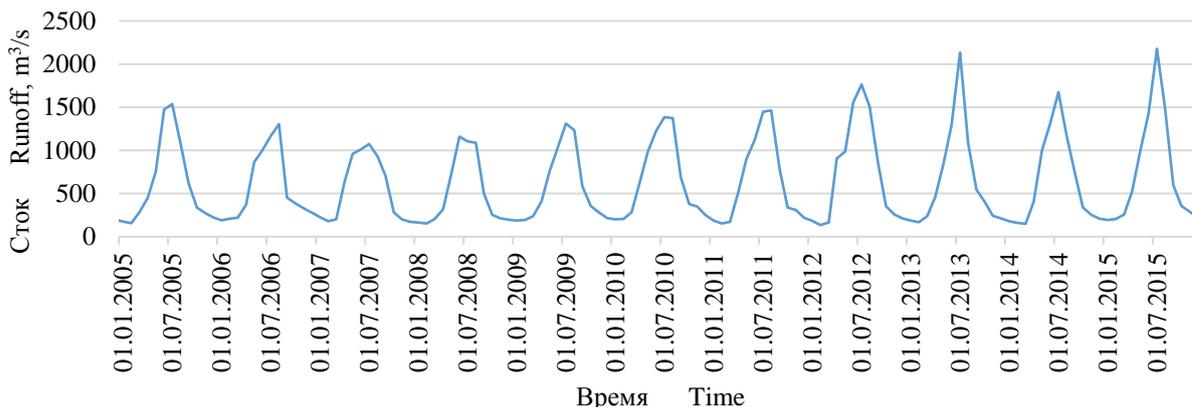


Рис.3. Сток реки Вахш 2005-2015 гг.
Fig. 3. Vakhsh River runoff for 2005-2015 years.

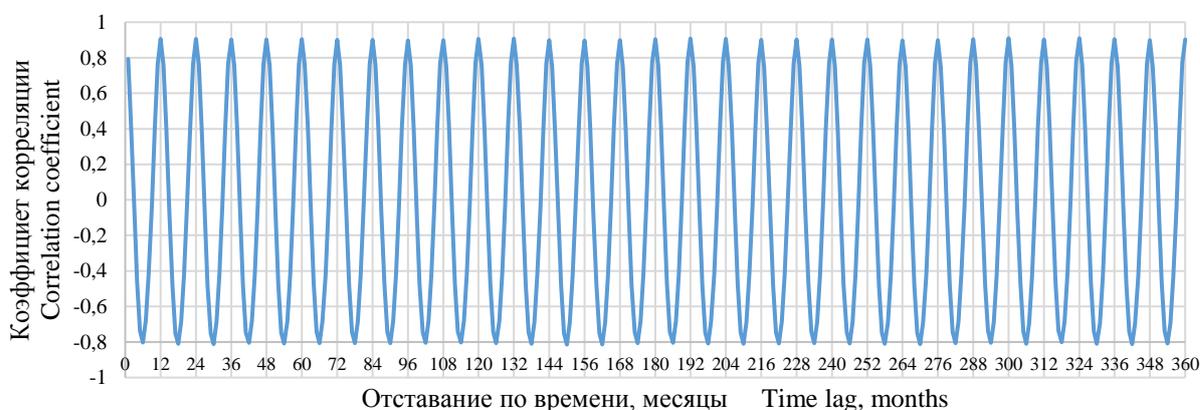


Рис.4. Коэффициент корреляции между стоком в определенный месяц и в предыдущие.
Fig. 4. Correlation coefficient between runoff in a certain month and in previous.

Примененный подход прогнозирования с помощью машинного обучения основан на построении моделей вида:

$$y_{t+h}^* = f(y_{t-w}, y_{t-w+6}, \dots, y_t), \quad (1)$$

где y_t – значение стока в месяц текущем месяце t ;

h – горизонт планирования в месяцах, кратный шести;

y_{t+h}^* – прогноз стока на h месяцев вперед;

w – ширина окна для выбора ретроспективных данных в месяцах, кратная шести.

Как было указано во введении, рассмотрено применение классического алгоритма kNN с построением более сложного пространства признаков. Рассмотрены следующие варианты:

- 1) использование исходных признаков (1);
- 2) использование логарифмированных значений признаков:

$$y_{t+h}^* = f(\log(y_{t-w}), \log(y_{t-w+6}), \dots, \log(y_t)); \quad (2)$$

- 3) использование полиномиальных признаков (2-й степени):

$$y_{t+h}^* = f(y_{t-w}, y_{t-w+6}, \dots, y_t, y_{t-w} \cdot y_{t-w}, y_{t-w} \cdot y_{t-w+6}, \dots, y_{t-w} \cdot y_t, \dots, y_t \cdot y_t); \quad (3)$$

- 4) использование полиномиальных логарифмированных признаков (2-й степени):

$$y_{t+h}^* = f(g_{t-w}, g_{t-w+6}, \dots, g_t, g_{t-w} \cdot g_{t-w}, g_{t-w} \cdot g_{t-w+6}, \dots, g_{t-w} \cdot g_t, \dots, g_t \cdot g_t), \quad (4)$$

$$g_j = \log(y_j), \quad i = t - w, t - w + 6, \dots, t.$$

Для исследования эффективности используемого подхода было проведено не только исследование влияния предложенного преобразования пространства признаков, но и сравнение с результатами других методов, которые

наиболее часто используют в работах по прогнозированию речного стока:

- авторегрессионные методы: AR, ARIMA;
- простейший метод – линейная регрессия (LR), но также с различными пространствами признаков;
- методы машинного обучения без использования метеорологических факторов на примере Random Forest [28] с подбором гипер-параметров;
- методы машинного обучения с использованием метеорологических факторов на примере многослойного персептрона (MLP – multilayer perceptron).

Для реализации последнего пункта были использованы доступные результаты наблюдений за следующими метеорологическими параметрами на метеостанциях Санглок, Яван и Лахш в Республике Таджикистан (за период 2002–2015 гг.):

- температура ($^{\circ}\text{C}$),
- влажность (%),
- уровень осадков (мм),
- уровень снежного покрова (мм).

Поскольку выполняется прогнозирование среднемесячных значений притока, метеорологические параметры усреднялись.

III. РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Вычислительные эксперименты проведены по следующему сценарию.

1. Сравнение моделей. При этом использовался подход кросс-валидации, выборка делилась пятью разными способами в соотношении 80 % на обучение, 20 % на валидацию. Итоговые метрики точности усреднялись по пяти валидационным подвыборкам. Был выбран горизонт планирования $h = 2$ года, и ширина окна $w = 10$ лет. Рассматривались все четыре указанные варианта формирования признаков.

2. Анализ влияния ширины окна на точность прогноза выбранной модели при различных горизонтах планирования. При этом выборка была разделена в пропорции 90 % для обучения модели, 10 % для тестирования. В тестовую часть попали более новые данные, в обучающую – более старые.

3. Анализ распределения ошибок лучшей выбранной модели при выбранной ширине окна.

Использовались реализации алгоритмов машинного обучения из библиотеки Scikit

Learn (scikit-learn.org). В результате подбора гипер-параметров алгоритмов были выбраны следующие: для kNN число $k = 20$, метрика расстояния – евклидово расстояние; для Random Forest число деревьев 40, максимальная глубина дерева 6.

Результаты показаны в Таблице 1, использованы средняя относительная ошибка по модулю (MAPE – mean absolute percentage error) и коэффициент детерминации R^2 .

Для следующих этапов исследования выбран алгоритм kNN с использованием полиномиальных логарифмированных признаков. Результаты анализа влияния ширины окна при различных горизонтах прогноза приведены в Таблице 2.

IV. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

A. Анализ полученных результатов

Из Таблицы 1 видно, что наименьшая ошибка получена при использовании kNN с полиномиальными логарифмированными признаками. Для Random Forest логарифмирование не оказало существенного влияния на точность, так как деревья решений не выполняют вычислений с признаками, а только подбирают для них пороги в правилах.

Также из Таблицы 1 видно, что использованием метеофакторов не повысило точность по сравнению с моделями машинного обучения, не использующими их. Это объясняется тем, что при долгосрочном прогнозировании существенное влияние оказывает изменение климата в рассматриваемой области Республики Таджикистан [23–24]. Кроме того, данных метеонаблюдений для исследуемого объекта намного меньше, чем данных наблюдений за стоком (14 лет против 89).

Из Таблицы 2 видно, что ширина окна 20 лет является наилучшим выбором. Хотя при этом не всегда получена минимальная ошибка, но всегда близкая к ней. Это говорит о том, что для построения прогноза необходимо учитывать данные за длительный период, алгоритм kNN находит наиболее подходящие текущим показателям фрагменты за весь период наблюдений и, исходя из них, формирует прогноз.

Также в работе проведен анализ распределения ошибок на тестовой выборке при использовании ширины окна 20 лет. Рисунки 5 и 6 показывают распределение ошибок при прогнозировании на 5 лет вперед. Видно, что большая часть ошибок находится в диапазоне от -20 до 20 %. Высокие ошибки в процентах

наблюдаются чаще всего в месяцы с малым стоком, как видно на Рисунке 7, из-за того, что величина ошибки *MAPE* делится на малую величину.

По абсолютным значениям стока большая часть ошибок попадает в диапазон от -350 до 250 м³/с. Большое число ошибок в сторону занижения связано с тем, что пиковые значения

притока, характерные для некоторых лет тестовой выборки в 2013 и 2015 гг. входят в восемь самых высоких значений за всю историю наблюдений. Рисунок 8 показывает результат наложения прогноза на истинные значения для тестовой выборки.

Таблица 1

Сравнение регрессионных моделей

Table 1

Comparison of Regression Models

Модель Model	Подход Approach	Пространство признаков Feature space	MAPE, %	R2
ARIMA	Статистический Statistical	исходное initial	22.92	0.87
S-ARIMA	Статистический Statistical	исходное initial	15.56	0.89
LR	Статистический Statistical	полином 2-й степени 2nd order polynomial	27.05	0.83
LR	Статистический Statistical	логарифмирование logarithmation	30.00	0.86
LR	Статистический Statistical	полином 2-й степени из лог. признаков logarithmation 2nd order polynomial	19.38	0.89
kNN	Машинное обучение Machine Learning	исходное initial	15.39	0.90
kNN	Машинное обучение Machine Learning	полином 2-й степени 2nd order polynomial	15.23	0.90
kNN	Машинное обучение Machine Learning	логарифмирование logarithmation	15.23	0.90
kNN	Машинное обучение Machine Learning	полином 2-й степени из лог. признаков logarithmation 2nd order polynomial	15.08	0.90
Random Forest	Машинное обучение Machine Learning	исходное initial	15.92	0.89
Random Forest	Машинное обучение Machine Learning	полином 2-й степени 2nd order polynomial	15.37	0.90
Random Forest	Машинное обучение Machine Learning	логарифмирование logarithmation	15.91	0.89
Random Forest	Машинное обучение Machine Learning	полином 2-й степени из лог. признаков logarithmation 2nd order polynomial	15.46	0.90
MLP	Машинное обучение + учет метеофакторов Machine Learning + accounting for meteorological factors	исходное + метеорологические факторы initial + meteorological factors	18.57	0.82

В. Значимость полученных результатов

В результате исследования на примере данной задачи показано, что логарифмирование входных признаков и последующее построение на их основе полинома второй степени повышает точность алгоритма kNN. Логарифмирование признаков является хорошо известным приемом предобработки данных для применения к ним машинного обучения, но использование полинома из входных признаков

принято использовать как этап предобработки только для линейной регрессии.

Поэтому отличие предложенного подхода заключается в объединении обоих способов и применении полученной трансформации признаков именно для алгоритма kNN. Это не усложняет сам алгоритм kNN, не снижает его интерпретируемость и не повышает риск переобучения.

Предлагаемый метод прогнозирования стока реки Вахш позволит решить задачу повышения эффективности планирования выработки электроэнергии при назначении оптимальных режимов работы каскада ГЭС, в том

числе уменьшить холостые сбросы и выполнять точное стратегическое планирование развития энергосистемы Республики Таджикистан.

Таблица 2

Влияние ширины окна w на точность прогноза

Table 2

Influence of the window width w on the prediction accuracy

Горизонт планирования h , лет Planning horizon h , years	Ширина окна w , лет Window width w , years				
	3	5	10	15	20
0.5	14.66	13.57	13.93	13.31	13.24
1	14.25	13.06	13.78	13.50	13.24
2	14.68	14.07	13.28	13.54	13.41
5	14.56	14.16	13.60	14.16	13.79
10	13.54	13.34	13.98	14.14	13.40

* в ячейках содержатся значения MAPE, выделены два минимальных значения для каждой строки
* cells contain MAPE, two minimum values are highlighted for each row

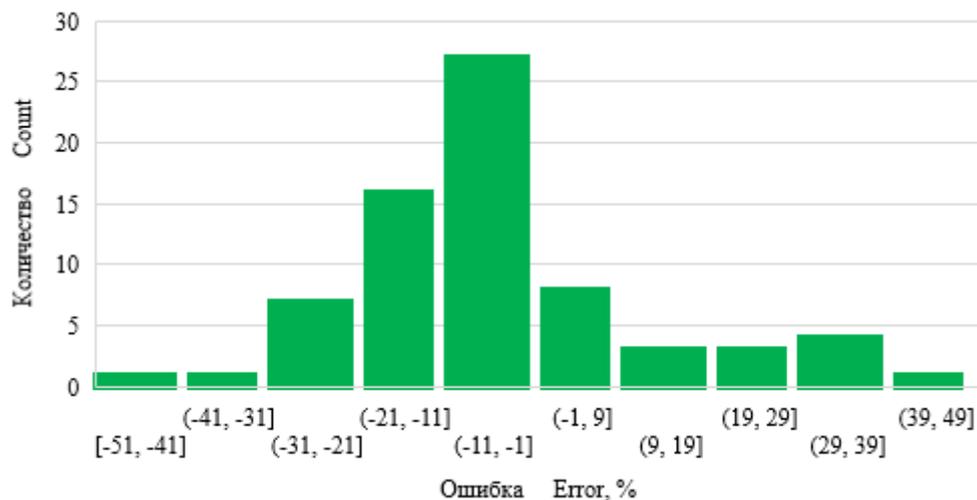


Рис.5. Распределение ошибок прогноза на 5 лет вперед, %.
Fig. 5. Distribution of forecast errors for 5 years ahead, %.

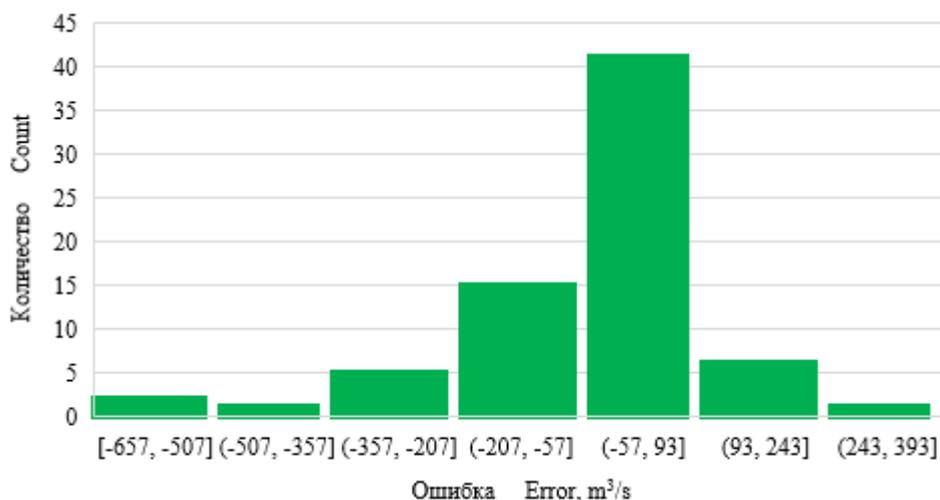


Рис.6. Распределение ошибок прогноза на 5 лет вперед, м³/с.
Fig. 6. Distribution of forecast errors for 5 years ahead, m³/s.

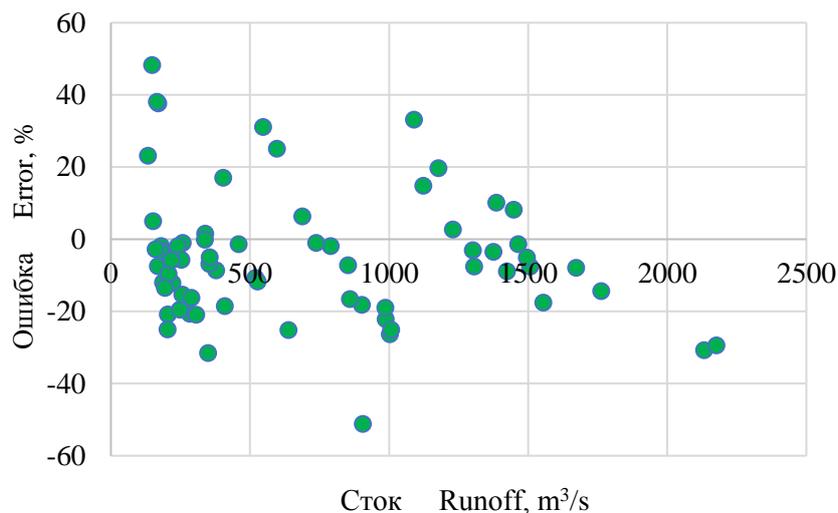


Рис.7. Распределение ошибок прогноза в зависимости от уровня стока.
Fig. 7. Distribution of forecast errors depending on runoff level.

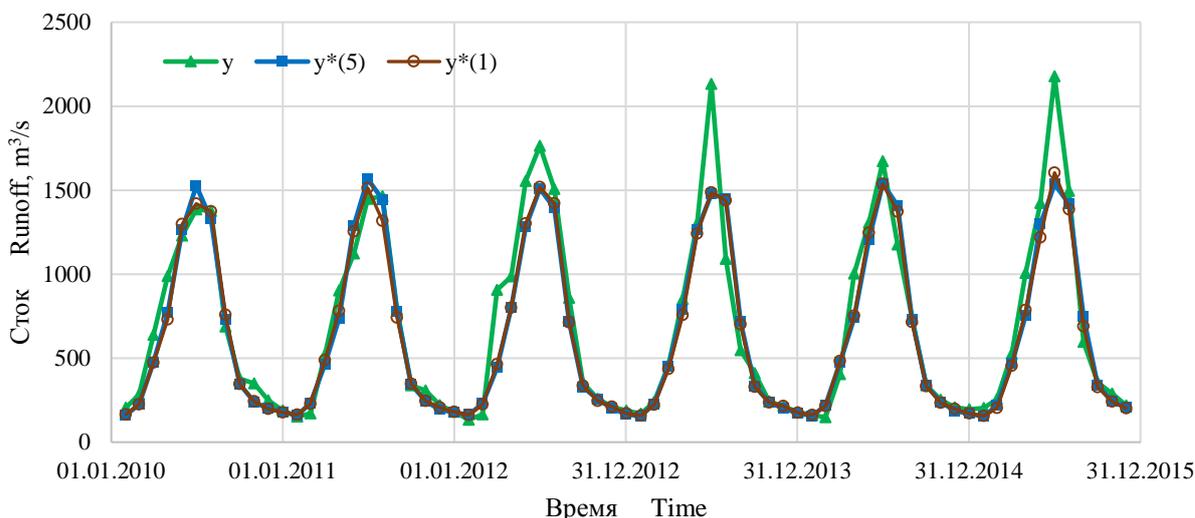


Рис.8. Сопоставление истинных данных y и прогнозов y^* на 1 год и 5 лет вперед.
Fig. 8. Comparison of true data y and forecasts y^* for 1 year and 5 years ahead.

ЗАКЛЮЧЕНИЕ

В работе выполнено построение модели долгосрочного прогнозирования стока реки Вахш. Для данного объекта долгосрочное прогнозирование с применением методов машинного обучения выполнено впервые. Вычислительные эксперименты выполнены на данных наблюдений за стоком реки с 1927 по 2015 гг. Предложена модифицированная непараметрическая модель на основе алгоритма k -ближайших соседей, новизна которой заключается в полиномиальном логарифмическом преобразовании пространства признаков, повышающем точность выбора подобных ретроспективных участков временного ряда. Полученная средняя по модулю ошибка прогноза на 1

год составила 13.24 % ($85 \text{ м}^3/\text{с}$), на 5 лет – 13.79 % ($90 \text{ м}^3/\text{с}$).

Применение моделей вида «черный ящик» позволяет получать долгосрочные прогнозы стока горных рек с сильно выраженной сезонностью, таких как река Вахш. Коэффициент детерминации составил 0,9, что соответствует результатам мирового уровня при использовании более сложных физико-математических моделей [6].

Показана важность использования многолетних наблюдений в задаче долгосрочного прогнозирования речного стока. Прогноз стока на 1–10 лет вперед следует вычислять на основании значений стока за предыдущие 20

лет, а для обучения прогнозной модели необходимы наблюдения как минимум за 50 лет.

В ходе дальнейших работ планируется выполнить анализ возможности повышения точности прогноза за счет учета в модели метеорологических наблюдений и создание программного продукта для его применения конечными пользователями.

ACKNOWLEDGEMENTS

Исследование выполнено при финансовой поддержке в рамках реализации программы развития НГТУ, научный проект №С22-15.

Литература (References)

- [1] Barkans J., Zicmane I. Electricity production by hydro power plants: possibilities of forecasting. *Latvian Journal of Physics and Technical Sciences*, 2004, no. 1, pp. 14-22.
- [2] International Energy Agency. *Data & Statistics*. 2020. Available at: <https://www.iea.org/> (accessed 21.01.2021)
- [3] Ekanayake P., Wickramasinghe L., Jayasinghe J.M.J.W., Rathnayake U. Regression-Based Prediction of Power Generation at Samanalawewa Hydropower Plant in Sri Lanka Using Machine Learning. *Mathematical Problems in Engineering*, 2021, vol. 2021, ID 4913824, pp. 1-12. doi: 10.1155/2021/4913824
- [4] Berga L. The Role of Hydropower in Climate Change Mitigation and Adaptation: A Review. *Engineering*, 2016, vol. 2, issue 3, pp. 313-318. doi: 10.1016/J.ENG.2016.03.004
- [5] Ghimire S., Yaseen Z.M., Farooque A.A., Deo R.C., Zhang J., Tao X. Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Scientific Reports*, 2021, vol. 11, article no. 17497, pp. 1-26. doi: 10.1038/s41598-021-96751-4
- [6] Motovilov U.G., Gelfan A.N. *Modeli formirovaniya stoka v zadachah gidrologii rechnyh bassejnov* [Models of Runoff Formation in River Basin Hydrology]. Moscow, 2018. 300 p.
- [7] Bates P. *Spatial patterns in catchment hydrology: observations and modelling*. Cambridge University Press, Cambridge, 2000. 404 p. doi: 10.1002/esp.378
- [8] Pyankov S.V., Shikhov A.N. *Geoinformacionnoe obespechenie modelirovaniya gidroloicheskikh processov I yavlenij* [Geoinformation support of modeling of hydrological processes and phenomena]. Perm, 2017. 148 p.
- [9] Neteler M., Bowman M. H., Landa M., Metz M. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software*, 2012, vol. 31, pp. 124-130. doi: 10.1016/J.ENVSOF.2011.11.014
- [10] Borsch S., Khristoforov A., Krovotyntsev V., Leontieva E., Simonov Y., Zatyagalova, V. A Basin Approach to a Hydrological Service Delivery System in the Amur River Basin. *Geosciences*, 2018, vol. 8, issue 3, pp. 1-16. doi: 10.3390/geosciences8030093
- [11] Xu C.-Y., Xiong L., Singh V. P. Black-Box Hydrological Models. In *Handbook of Hydrometeorological Ensemble Forecasting*, Springer Berlin Heidelberg, 2017, pp. 1-48. doi: 10.1007/978-3-642-40457-3_21-1
- [12] Nair J. P., Vijaya M. S. Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 1747-1753. doi: 10.1109/ICAIS50930.2021.9395832
- [13] Cover T., Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967, vol. 13, no. 1, pp. 21-27. doi: 10.1109/TIT.1967.1053964
- [14] Maillo J., García S., Luengo J., Herrera F., Triguero I. Fast and Scalable Approaches to Accelerate the Fuzzy k-Nearest Neighbors Classifier for Big Data. *IEEE Transactions on Fuzzy Systems*, 2020, vol. 28, no. 5, pp. 874-886. doi: 10.1109/TFUZZ.2019.2936356
- [15] Bergstra J., Bengio Y. Random search for hyperparameter optimization. *Journal of Machine Learning Research*, 2012, vol. 13, pp. 281-305.
- [16] Zhang S., Li X., Zong M., Zhu X., Wang R. Efficient kNN Classification With Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, vol. 29, no. 5, pp. 1774-1785. doi: 10.1109/TNNLS.2017.2673241
- [17] Nguyen B., Morell C., de Baets B. Large-scale distance metric learning for k-nearest neighbors regression. *Neurocomputing*, 2016, vol. 214, pp. 805-814. doi: 10.1016/j.neucom.2016.07.005
- [18] Paredes R., Vidal E. Learning weighted metrics to minimize nearest-neighbor classification error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, vol. 28, no. 7, pp. 1100-1110. doi: 10.1109/TPAMI.2006.145
- [19] Biswas N., Chakraborty S., Mullick S. S., Das S. A parameter independent fuzzy weighted k-Nearest neighbor classifier. *Pattern Recognition Letters*, 2018, vol. 101, pp. 80-87. doi: 10.1016/J.PATREC.2017.11.003
- [20] Hou W., Li D., Xu C., Zhang H., Li T., An Advanced k Nearest Neighbor Classification Algorithm Based on KD-tree. *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, 2018, pp. 902-905. doi: 10.1109/IICSPI.2018.8690508
- [21] Lu D., Ning Q., Zang J. Improved KNN algorithm based on BP neural network decision making. *Journal of Computer Applications*, 2017, vol. 37, no. 2, pp. 65-67, 2017.

- [22] Laldjebaev M., Isaev, R., Saukhimov, A. Renewable energy in Central Asia: An overview of potentials, deployment, outlook, and barriers. *Energy Reports*, 2021, vol. 7, pp. 3125-3136. doi: 10.1016/j.egyр.2021.05.014
- [23] Xenarios S., Laldjebaev M., Shenhav R. Agricultural water and energy management in Tajikistan: a new opportunity. *International Journal of Water Resources Development*, 2021. vol. 37, no. 1, pp. 118-136. doi: 10.1080/07900627.2019.1642185
- [24] Gulakhmadov A., Chen X., Gulakhmadov N., Liu T., Anjum M. N., Rizwan M. Simulation of the Potential Impacts of Projected Climate Change on Streamflow in the Vakhsh River Basin in Central Asia under CMIP5 RCP Scenarios. *Water*, 2020, vol. 12, no. 5, pp. 1-34. doi: 10.3390/w12051426
- [25] Gulakhmadov A., Chen X., Gulakhmadov M., Kobuliev Z., Gulakhmadov, N., Peng J., Liu T. Evaluation of the CRU TS3. 1, APHRODITE_V1101, and CFSR Datasets in Assessing Water Balance Components in the Upper Vakhsh River Basin in Central Asia. *Atmosphere*, vol. 12, no. 5, pp. 1-40. doi: 10.3390/atmos12101334
- [26] Jain G., Mallick, B. A Study of Time Series Models ARIMA and ETS. *International Journal of Modern Education and Computer Science*, 2017, vol. 9, no. 4, pp. 57-63. doi: 10.5815/ijmecs.2017.04.07
- [27] Valipour M. Long-term runoff study using SARIMA and ARIMA models in the United States. *Meteorological Applications*, 2015, vol. 22, iss. 3, pp. 592-598. doi: 10.1002/met.1491
- [28] Breiman L. Random Forests. *Machine Learning*, 2001, vol. 45, iss. 1, pp. 5-32. doi: 10.1023/A:1010933404324

Сведения об авторах.



Матренин Павел Викторович, кандидат технических наук. Доцент кафедры «Системы электроснабжения предприятий» Новосибирского государственного технического университета. Область научных интересов: системный анализ, методы машинного обучения, оптимизация и управление в электроэнергетике.
E-mail: matrenin.2012@corp.nstu.ru



Кирьянова Наталья Геннадьевна, ассистент кафедры «Автоматизированные электроэнергетические системы» Новосибирского государственного технического университета. Область научных интересов: моделирование и оптимизация электроэнергетических систем, системы накопления энергии
E-mail: kiryanova-ng@ya.ru



Сафаралиев Муродбек Холназарович, инженер исследователь кафедры «Автоматизированные электрические системы» Уральского федерального университета имени первого Президента России Б.Н. Ельцина. Область научных интересов: модели планирования, управления и развития энергетических систем.
E-mail: murodbek_03@mail.ru



Султонов Шерхон Муртазокулович, кандидат технических наук, доцент кафедры «Электрические станции» Таджикского технического Университета имени академика М.С. Осими. Область научных интересов: оптимизация управления режимами гидроэлектростанций.
E-mail: Sultonzoda@ttu.tj