**Оценка содержания озона в атмосфере на Байкальской природной территории по наземным данным и по данным реанализа**

Сметанина А.М.1,2, Громов С.А.2,1, Оболкин В.А.3, Ходжер Т.В.3, Хуриганова О.И.3

1) Институт географии РАН

2) Институт глобального климата и экологии им. Ю.А. Израэля

3) Лимнологический институт СО РАН

Озон в тропосфере является токсичным газом, оказывающим негативное влияние на здоровье человека. На основании метеорологических и других параметров реанализа была построена модель машинного обучения для прогноза концентрации озона на станции Листвянка в Байкальском регионе. Обучение и верификация модели проводились на данных измерений автоматического газоанализатора, осуществляющего мониторинг содержания озона на станции Листвянка. В качестве моделей машинного обучения мы использовали случайный лес и бустинг. Среднее абсолютное отклонение реанализа составляет более 16 единиц на миллиард, а среднее процентное абсолютное отношение составило 80%. Для смоделированных значений отклонение составило 6 единиц на миллиард и 29%. соответственно. Наиболее важными параметрами в моделировании явились номер месяца, температура и растительность. С помощью полученной модели машинного обучения и параметров реанализа было смоделировано содержание озона на 2017-2022 годы.

Ключевые слова: тропосферный озон, реанализ, машинное обучение

**Ozone concentration assessment from reanalysis and ground-based measurements in Baikal region**

Smetanina A.M.1,2, Gromov S.A.2,1, Obolkin V.A.3, Khodzher T.V.3, Khuriganova O.I.3

1) Institute of Geography RAS

2) Yu.A. Izrael Institute of Global Climate and Ecology

3) Limnological Institute SB RAS

Tropospheric ozone is a toxic gas causing negative impact on human health. To predict ozone concentrations at Listvyanka monitoring station in Baikal region we used machine learning model. Features for the model were taken from ERA5 reanalysis. The model was trained and verified on automatic monitor ground-based measurements of ozone concentrations. We used random forest and boosting machine learning models. Reanalysis data of ozone concentrations mean absolute error exceeds 16 ppb. Mean percentage error is about 80 %. The errors of the modelled values of ozone concentrations equal to 6 ppb and 29% respectively. The most important features in modelling are month number, high vegetation leaf area index and temperature. Ozone concentrations for 2017 – 2022 year are predicted using the trained model.

Key words: tropospheric ozone, reanalysis, machine learning

**Введение**

Озон в тропосфере является токсичным газом, оказывающим негативное влияние на растительность и на здоровье человека. Мониторинг озона в тропосфере ведется различными международными сетями наблюдений. Существует международный проект, посвященный оценке содержания озона в тропосфере (Tropospheric Ozone Assessment Report, TOAR) (<https://igacproject.org/activities/TOAR>).

Озон – фотохимически активный газ, является одной из составляющих фотохимического смога. На его содержание в атмосфере влияет ряд факторов, среди которых солнечное излучение, температура, влажность, ветер и др. Большую роль в формировании и стоках озона играют антропогенные и природные источники прекурсоров озона, таких как оксиды серы и азота, летучие органические соединения, фитонциды (Белан, 1996). Также на содержание озона может влиять наличие снежного покрова (Белан, 2018).

В последнее время все чаще используются технологии машинного обучения для оценки и прогноза содержания озона в приземном слое воздуха для оценки загрязнения воздуха, например, в городах (Balamurugan, 2022; Carbo-Bustinza, 2022 и др.). В качестве входных параметров в моделях машинного обучения используются как прекурсоры озона, так и метеопараметры, а также другие характеристики.

Целью данной статьи является построение модели машинного обучения, которая на основании входных параметров (признаков) будет рассчитывать концентрацию приземного озона на станции Листвянка Иркутской области. В качестве признаков, подаваемых на вход модели, были взяты данные реанализа ERA5 (Copernicus…, 2017). В качестве целевой функции мы использовали данные измерений концентраций озона автоматическим газоанализатором на станции Листвянка.

**Материалы и методы**

Станция Листвянка расположена недалеко от поселка Листвянка на берегу озера Байкал, на высоте 650 м. Рядом находится обсерватория института солнечно-земной физики СО РАН. Непосредственно вокруг станции нет антропогенных источников, однако рядом расположен одноименный поселок. Также иногда прямо над станцией проходит шлейф выбросов ТЭЦ, который также оказывает непосредственное влияние на состав воздуха на станции (Оболкин, 2019). Измерения содержания озона осуществляются с помощью автоматического газоанализатора. Временное разрешение данных о концентрации озона составляет 5 минут.

Мы использовали данные измерений автоматического газоанализатора за 2015 и 2016 годы. В качестве обучающего массива данных был взят период с начала 2015 до середины 2016 года. Вторая половина 2016 года использовалась для валидации модели.

В реанализе ERA5 содержатся данные с пространственным разрешением 0.25̊×0.25̊. Для решения задачи мы использовали данные с временным разрешением 6 часов. В реанализе доступен широкий выбор параметров, в том числе, концентрация озона у поверхности. В качестве признаков для моделей МО мы выбирали параметры, которые в реальности могут влиять на формирование озона у поверхности земли и используются для моделирования содержания озона (Белан, 2010; Cao, 2023). В качестве признаков модели мы рассматривали температуру, давление, скорость и направление ветра, растительность, солнечное излучение и другие, а также содержание озона у поверхности земли по данным реанализа ERA5. Прекурсоры озона, такие как, например, оксиды азота, в качестве признаков не рассматривались.

На рисунке 1 приведены временные ряды концентраций озона на станции Листвянка по данным реанализа ERA5 и по данным автоматического газоанализатора, осредненным с шестичасовым разрешением.

Рисунок 1. Концентрации озона у поверхности земли по данным автоматического газоанализатора (АМ), и по данным реанализа (ERA5)

Из рисунка 1 видно, что в 2015 и половине 2016 годов данные реанализа в целом соответствуют ходу содержания озона, однако имеет место значительная переоценка в холодные месяцы с ноября по март, а также амплитуда колебаний данных реанализа существенно ниже разброса реальных концентраций, даже с учетом осреднения по времени.

Для построения моделей машинного обучения мы использовали случайный лес (RandomForestRegressor библиотеки scikit learn) и бустинг (библиотека CatBoost). Случайный лес представляет собой ансамбль решающих деревьев. Бустинг представляет собой итерационную модель, где каждая следующая модель пытается уменьшить ошибку предыдущей. Подбор гиперпараметров моделей мы проводили с помощью библиотеки Optuna. Ниже приведен список оптимальных гиперпараметров для CatBoost и Random Forest.

Таблица 1. Значения оптимальных гиперпараметров моделей машинного обучения

**Результаты**

Случайный лес и бустинг показали очень схожие по точности результаты. Также оценка важности признаков в итоговой модели оказалась почти одинакова. Однако в процессе поиска оптимальных признаков в модели бустинга наибольший вклад имели физически объяснимые величины. Поэтому в качестве рабочей модели в итоге был выбран бустинг.

При сравнении моделей мы сравнивали среднеквадратичную ошибку (mean squared error, MSE), среднюю абсолютную ошибку (mean absolute error, MAE), которая менее чувствительна к случайным завышенным или заниженным значениям, а также среднее абсолютное процентное отклонение (mean absolute percentage error, MAPE).

В реанализе ERA5 озон является расчетной величиной (а не получаемой по каким-либо измерениям). Оказалось, что при моделировании, подаваемые на вход модели значения концентраций озона из реанализа ухудшают точность моделирования. Наилучшие результаты (наименьшую ошибку MAE) обе модели дают при использовании следующих (среди рассмотренных нами) признаков: номер месяца (month), температура (t), индекс площади поверхности листьев высокой растительности (high vegetation leaf area index, lai\_hv), относительная влажность (r), нисходящее солнечное излучение (ssrd), глубина снега (snow depth), u- и v- компоненты ветра (u, v), а также скорость ветра (w\_spd) и направление (wa, отношение u- и v- компонент). Индекс площади поверхности листьев высокой растительности оказался значительно более важным, чем тот же параметр для невысокой растительности.

В таблице 2 приведены ошибки моделей на тестовой выборке.

Таблица 2. Ошибки моделей машинного обучения и данных реанализа (ERA5) на тестовой выборке (вторая половина 2016 года)

Средняя абсолютная ошибка MAE ниже, чем MSE, что говорит о том, что модельные значения недостаточно хорошо оценивают максимумы и минимумы исходного временного ряда.

На рисунке 2а приведены результаты моделирования на валидационном периоде (вторая половина 2016 года). Видно, что значения концентраций озона из реанализа (ERA5) существенно завышены сравнению с реальными измерениями (АМ) как и в случае с холодными месяцами 2015 года. Однако и бустинг, и случайный лес показали хороший результат. И провалы, и пики хорошо моделируются, хотя модельные значения недостаточно отражают размах реальных значений.

На рисунке 2б приведен вклад обучающих признаков в моделях случайного леса и бустинга. Видно, что в обеих моделях значимость признаков одинакова в том смысле, что наиболее важным признаком оказался номер месяца, потом по значимости идет температура, индекс площади поверхности листьев, относительная влажность, давление, солнечное излучение, глубина снега, составляющие скорости ветра.

Рисунок 2. а) Прогноз моделей на валидационном периоде (автоматический газоанализатор АМ, бустинг CAT, случайный лес RF, реанализ ERA5); б) вклад признаков

Глубина снега вносит заметный вклад в обеих моделях, тогда такой признак как индекс наличия снега (есть снег или нет) при моделировании практически не внес никакого вклада, поэтому был исключен.

Информация о ветре уменьшает ошибку прогноза, однако не является решающей. В аналогичных исследованиях по озону данные о ветре вносят разный вклад в результирующую модель. Существуют станции, где содержание озона по результатам машинного моделирования в наибольшей степени определяется ветром (Carbo-Bustinza, 2022).

На рисунке 3 приведен среднемесячный прогноз концентраций озона на станции Листвянка построенной модели бустинга на 2017 – 2022 годы, построенный на основе признаков из реанализа, перечисленных выше. Также приведены оценки концентраций озона тоже по данным реанализа ERA5. И там и там мы видим выраженные колебания годового хода. Помимо различия в амплитудах годового хода концентраций озона имеет место также смещение пиков.

Рисунок 3. Прогноз концентраций озона на станции Листвянка по данным реанализа

Выводы

В целом, реанализ ERA5 может быть использован для оценки среднемесячных значений содержания озона у поверхности земли в Листвянке почти во все месяцы за исключением осени*.* В холодные месяцы 2015 и 2016 годов значения реанализа существенно завышены по сравнению с реальными измерениями. Амплитуда колебаний концентраций в реанализе практически не отражает реальный разброс. После обработки данных реанализа с помощью моделей машинного обучения, ошибка существенно снижается, намного лучше моделируется амплитуда разброса концентраций озона. Прогнозные модельные значения на валидационной выборке имеют среднюю абсолютную ошибку 6 частиц на миллиард, причем следует отметить, что эти данные получены исключительно на данных самого реанализа, без каких-либо локальных измерений. Возможно, при наличии локальных измерений, точность модели могла бы быть еще выше. При прогнозировании содержания озона на длительный срок необходима валидация на локальных измерениях

Работа выполнена по Госзаданию № АААА-А19-119022190168-8.

**Литература**

**1. Белан Б. Д**. Проблема тропосферного озона и некоторые результаты его измерений . // Оптика атмосферы и океана. 1996. Т. 9. № 09. С. 1184-1213

**2. Б. Д. Белан ; под ред. В. А. Погодаева** Озон в тропосфере // Российская акад. наук, Сибирское отд-ние, Ин-т оптики атмосферы им. В. Е. Зуева СО РАН. - Томск : Изд-во Ин-та оптики атмосферы СО РАН. 2010. - 487 с. : ил., табл.; 20 см.; ISBN 978-5-94458-117-4

**3. Белан Б. Д., Савкин Д. Е., Толмачев Г. Н.** Исследование связи снежного покрова и концентрации озона в приземном слое воздуха в районе г. Томска . // Оптика атмосферы и океана. 2018. Т. 31. № 08. С. 665–669. DOI: 10.15372/AOO20180811.

**4. Оболкин В.А., Шаманский Ю.В., Ходжер Т.В., Фалиц А.В.**, Мезомасштабные процессы переноса атмосферных загрязнений в районе южного Байкала // Океанологические исследования, 2019, Том 47, No 3, С. 104–113 DOI: 10.29006/1564-2291.JOR-2019.47(3).9

**5. Cao, Y., Zhao, X., Su, D., Cheng, X., Ren, H.**  A Machine-Learning-Based Classification Method for Meteorological Conditions of Ozone Pollution. // Aerosol Air Qual. Res. 23, 2023, 220239. https://doi.org/10.4209/aaqr.220239

**6. Carbo-Bustinza, N., Belmonte, M., Jimenez, V. *et al.*** A machine learning approach to analyse ozone concentration in metropolitan area of Lima, Peru. // *Sci Rep* **12**, 22084 (2022). https://doi.org/10.1038/s41598-022-26575-3

**7. Copernicus Climate Change Service (C3S)** ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS), date of access. 05/08/2023 <https://cds.climate.copernicus.eu/cdsapp#!/home>

**8. Pan, Q., Harrou, F. & Sun, Y.** A comparison of machine learning methods for ozone pollution prediction.// *J Big Data* **10**, 63 (2023). <https://doi.org/10.1186/s40537-023-00748-x>

**9. Balamurugan, V., Balamurugan, V. & Chen, J**. Importance of ozone precursors information in modelling urban surface ozone variability using machine learning algorithm. // *Sci Rep* **12**, 5646 (2022). https://doi.org/10.1038/s41598-022-09619-6