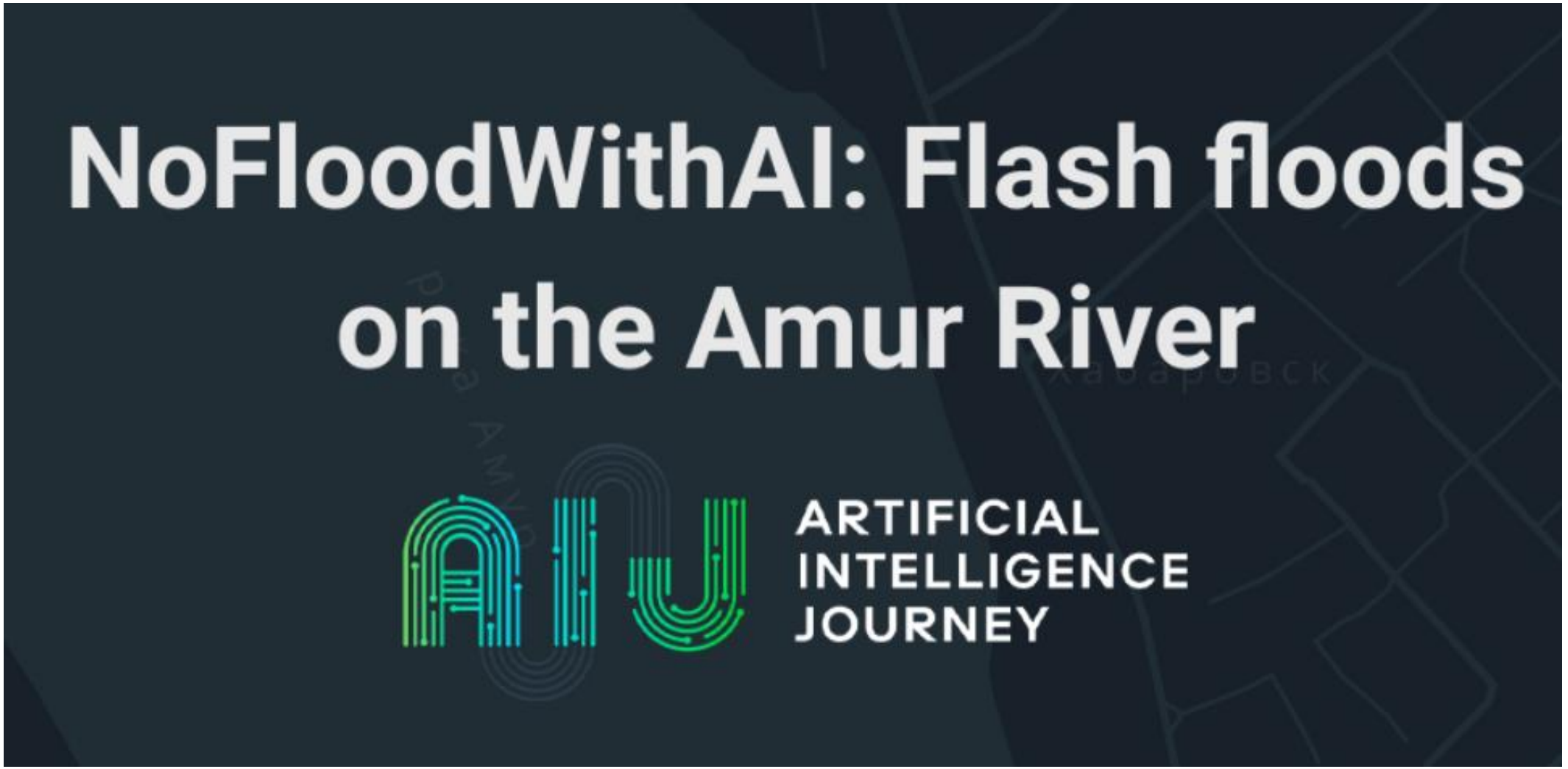


Модель прогнозирования уровня воды на реке Амур

Раевский Д.Н. nofirma2010@mail.ru RaevskyDN

Краткое описание модели

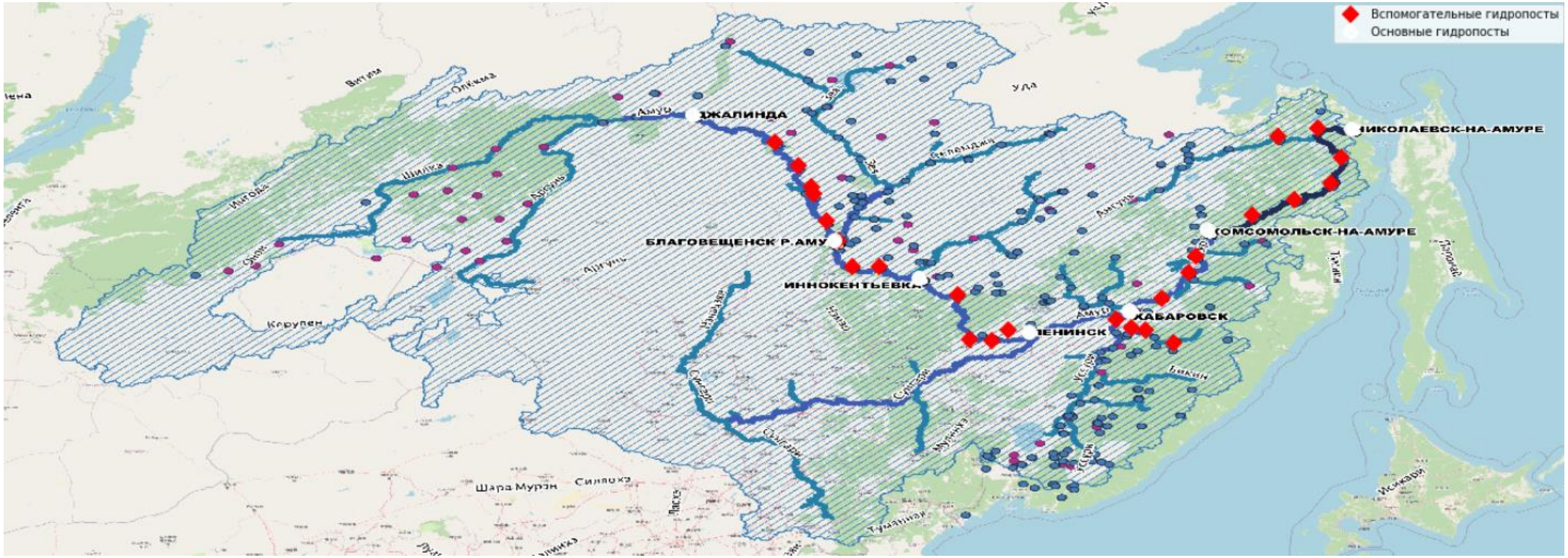
Модель: нейросеть (FC+LSTM+Conv)

Функция потерь: усредненная по гидропостам среднеквадратическая ошибка

Входная размерность: 10x1518

Выходная размерность: 10x33

Общий подход к построению модели



Используются данные сразу с 33 гидропостов, из них 8 – целевые.

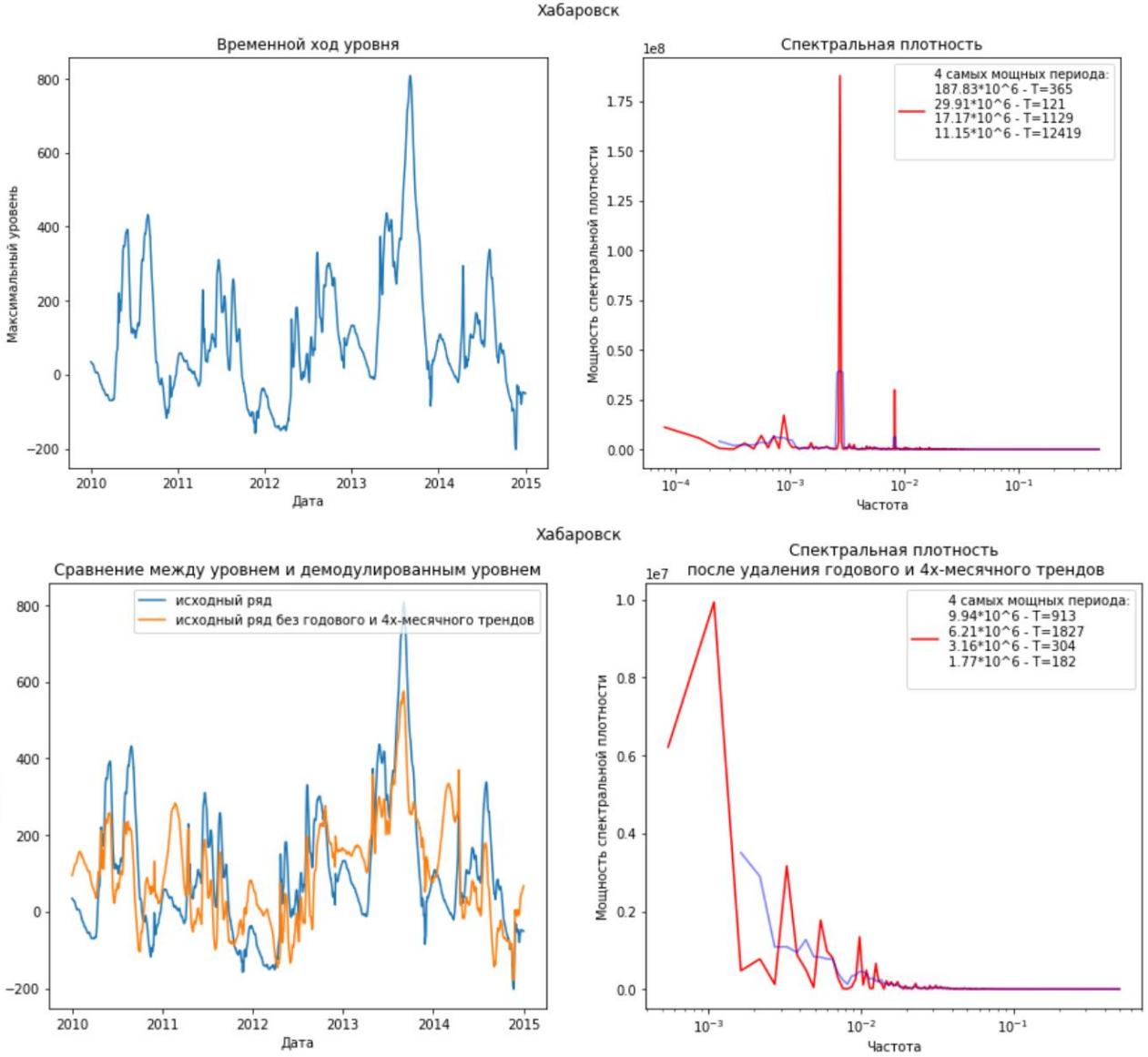
Уровень предсказывается сразу для 33 гидропостов.
Такой подход позволяет учесть гораздо больше информации по всему бассейну, а за счет высокой корреляции между уровнями различных гидропостов модель, предсказывающая уровень сразу в 33 точках, обучается качественнее, так как эта зависимость сохраняется.

Используемые признаки с каждой станции

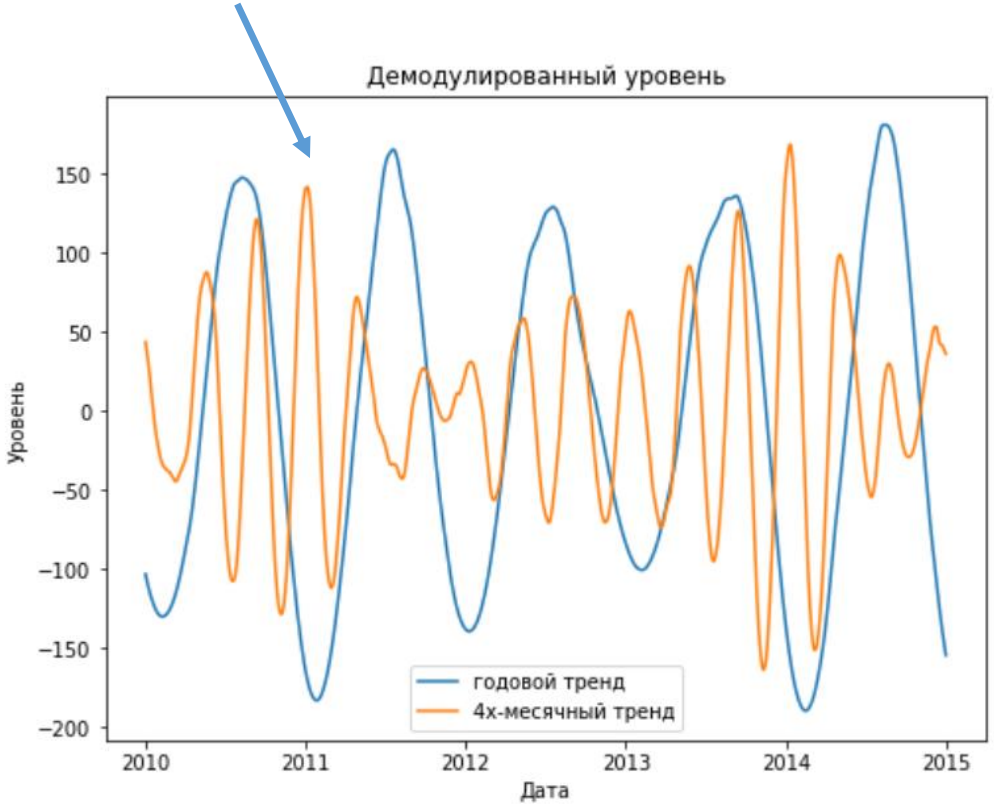
- **Значения максимального уровня воды 10,20,30 дней назад** – значения уровня и его динамика – один из важнейших признаков в этой модели, изменения в уровне за последнее время несут в себе большую информацию о гидрофизических- и метео-процессах, протекавших ранее, которые могут оказывать влияние на изменение текущего уровня (особенно в осенне-зимний период);
- **демодулированный уровень воды 10 дней назад для периодов 365, 121 дней** – для учета цикличности (подробнее на следующем слайде) ;
- **максимальный уровень воды год назад и максимальный уровень воды год назад, усредненный за три дня** – для учета локальных особенностей конкретного гидропоста;
- **метео-данные за текущий день (скорость ветра, направление ветра, атмосферное давление, сума осадков (мм), относительная влажность воздуха, макс. температура воздуха за день, мин. температура воздуха за день);**
- **метео-данные за вчерашний день (скорость ветра, направление ветра, атмосферное давление, сума осадков (мм), относительная влажность воздуха, макс. температура воздуха за день, мин. температура воздуха за день);**
- **метео-данные 10 дней назад (скорость ветра, направление ветра, атмосферное давление, сума осадков (мм), относительная влажность воздуха, макс. температура воздуха за день, мин. температура воздуха за день, максимальная скорость ветра, температура почвы)** – для учета более точных исторических данных;
- **средняя сумма осадков (мм) и средняя относительная влажность за последние 7 дней;**
- **средняя сумма осадков (мм) и средняя относительная влажность за последние 45 дней** – особенно важные признаки для весеннего и зимнего периодов;
- **количество дней без осадков;**
- **Sin,cos преобразования от текущей дня в году** – для учета цикличности.

Метод частотно-фазовой демодуляции

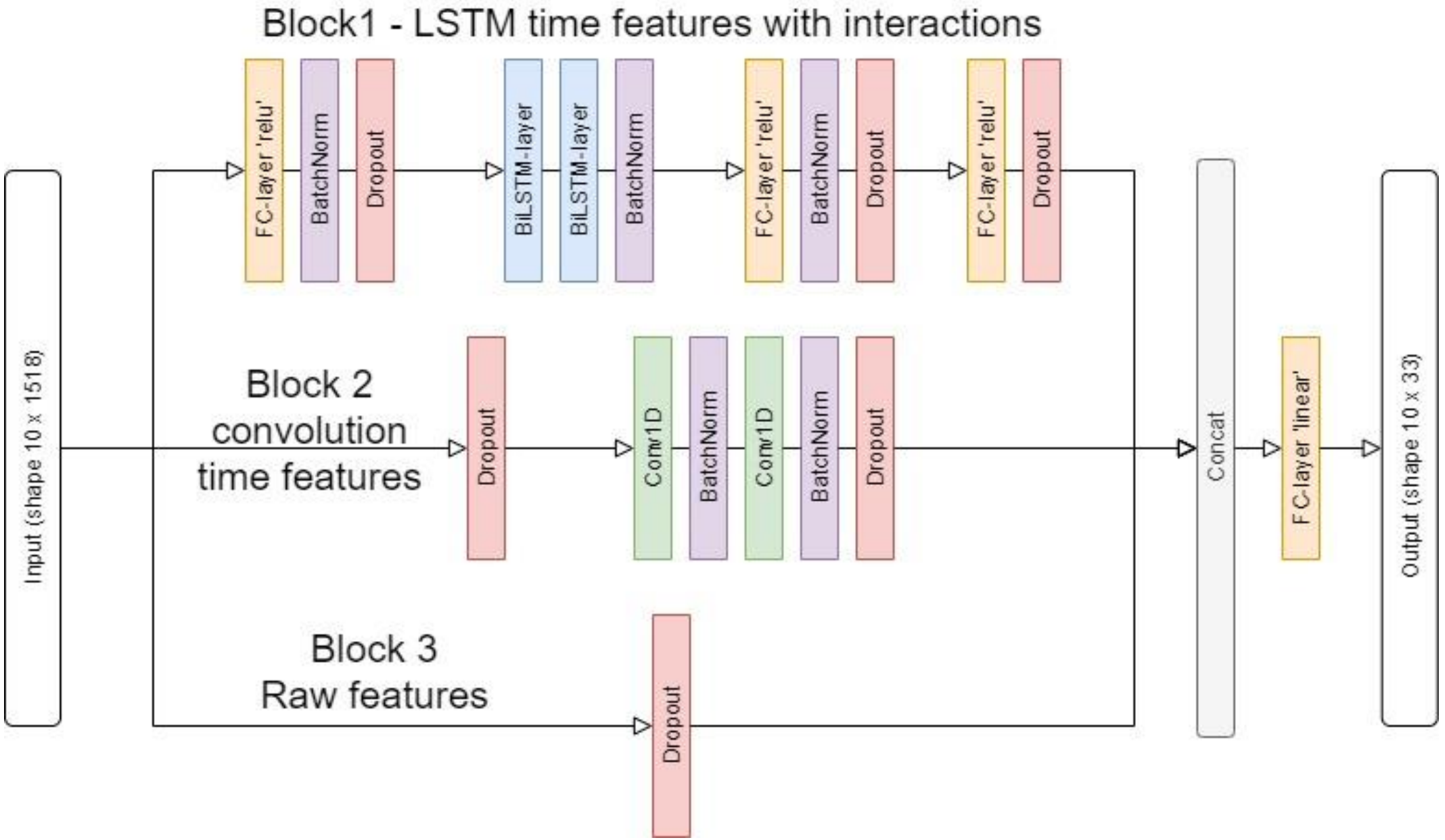
Метод частотно-фазовой демодуляции позволяет выделить контрольные периодические составляющие. Описание метода есть в [1], код приведен в репозитории. Рассмотрим на примере Хабаровска.



Эти признаки используются в модели



Архитектура нейросети



Обоснование выбора архитектуры

Такая архитектура во многом была обусловлена после прочтения некоторых статей по прогнозированию уровня воды, используемых физических и статистических характеристик (ниже приведены два примера).
Например:

1) метод соответственных уровней (см. [2]) – линейная комбинация между уровнями с разных станций, основан на расчетных временах добегания с разных станций. Если где-то есть такая зависимость, архитектура должна ее уловить. Вместе с **LSTM+conv features** этот учет возможен и для прогноза на 10 суток.

2) Индекс предшествующих осадков (см. [3]).

Линейная комбинация из запаса воды в снежном покрове, слое воды и среднее в скользящем окне по количеству осадков за сутки.

Такие признаки будут уловлены посредством комбинации **LSTM+raw features**.

$$m_{t_d} = (S+U)K^t + \sum_{J=0}^{j=t} x_{t-j}(K_{t-j})^j,$$

$$K_{t-j} = K_0 \exp[-c\theta_{t-j}], \qquad (1)$$

где S – запас воды в снежном покрове в конце зимы; U – слой воды, заполняющий водоудерживающую емкость бассейна в конце зимы, определяемый по эмпирической зависимости от показателя осеннего увлажнения; K_0, c – коэффициенты; x_{t-i} – количество осадков за сутки $t-i$, K_{t-i} – коэффициент, зависящий от среднесуточной температуры воздуха θ.

Всего было апробировано около 50 различных вариаций
Архитектур, в итоге была выбрана наилучшая по валидации и разницей в метриках между трейн/тест (с минимальным переобучением.

Использованные функции потерь

1) Среднеквадратическая ошибка (СКО) $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2 \to \min, \text{n} - \text{количество наблюдений}$

2) Модуль ошибки $\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y_i}| \to \min, \text{n} - \text{количество наблюдений}$

3) Усредненная по гидропостам СКО $\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(y_i-\hat{y_i})^2} \to \min, K - \text{количество гидропостов}$

4) Усредненный по гидропостам модуль ошибки $\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}|y_i-\hat{y_i}| \to \min$

5) Штрафующий за превышение СКО: $\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}f(y_i,\hat{y_i})}\, f(y_i,\hat{y_i}) = \begin{cases} c(y_i-\hat{y_i})^2, \hat{y_i} < y_i \\ \frac{1}{c}(y_i-\hat{y_i})^2, \hat{y_i} \geq y_i \end{cases}, c > 1$

В итоге был выбран вариант 3), т.к. лучше сохранялась корреляция между станциями, и модуль ошибки на целевых станциях была меньше. Однако возможно и использование функционала 5) – при увеличении параметра с модель будет чаще завышать относительно реального уровня, и точность определения опасных отметок уровня будет выше.

Обработка метео данных и использование прогнозных моделей

Инференс на исторических данных

При тестировании на исторических периодах метео данные используются из источника АСИОРИ.

Чтобы получить значения в точках гидропостах используется метод кригинга по трем ближайшим соседям (тестировались и другие более простые методы: ближайшие соседи, линейная аппроксимация, сплайны).

Если на какой-либо метеостанции по конкретному признаку за конкретный день нет данных или их качество сомнительно, то данные с этой метеостанции не учитываются.

Инференс при прогнозе на 10 дней вперед

При инференсе на 10 дней вперед используются сервисы с прогнозом погоды и исторические метео данные, обработка которых ведется

Прогноз погоды используются во всех гидропостах, информация с которых используется в модели (33). Сейчас подключены два сервиса, прогнозирующих погоду в точке:

1) Яндекс.погода

Яндекс Погода

2) Openweather

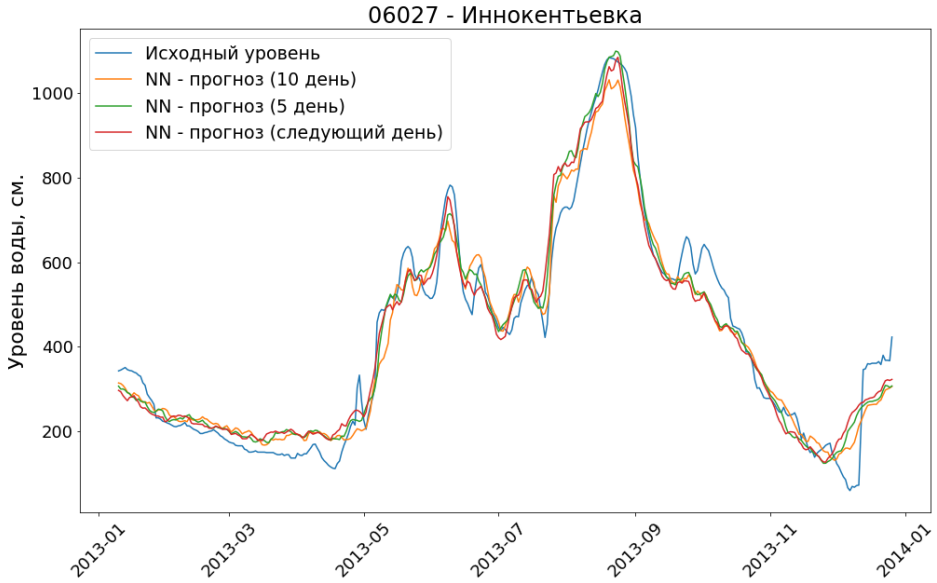
OpenWeather

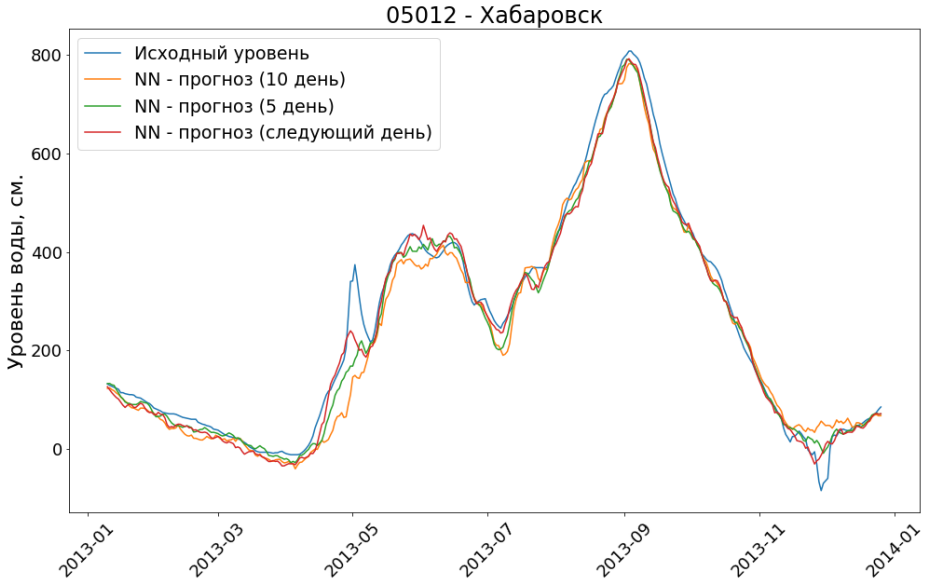
Результаты прогноза погоды объединяются по **пессимистическому сценарию**: выбирается максимальное количество осадков и влажности, мин. и макс. температуры воздуха.

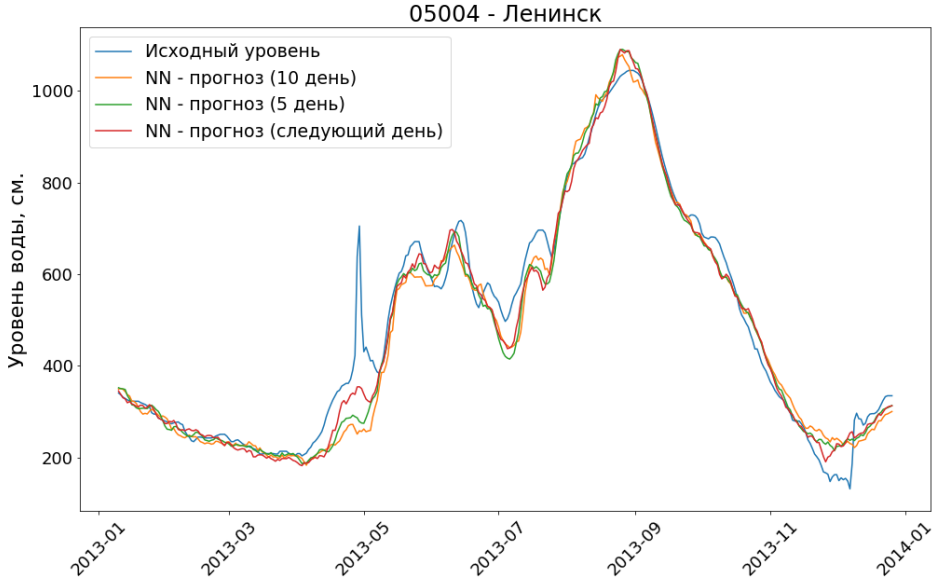
Результаты при различных сроках прогноза на примере 2013 г

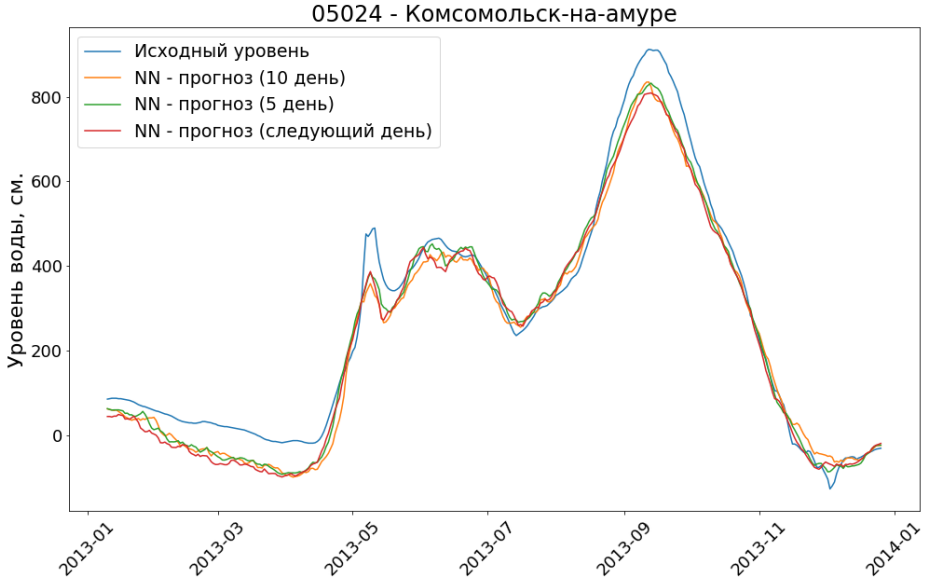
Обучение: все данные до 2013-01-01

Тест: 2013-01-01 – 2014-01-01





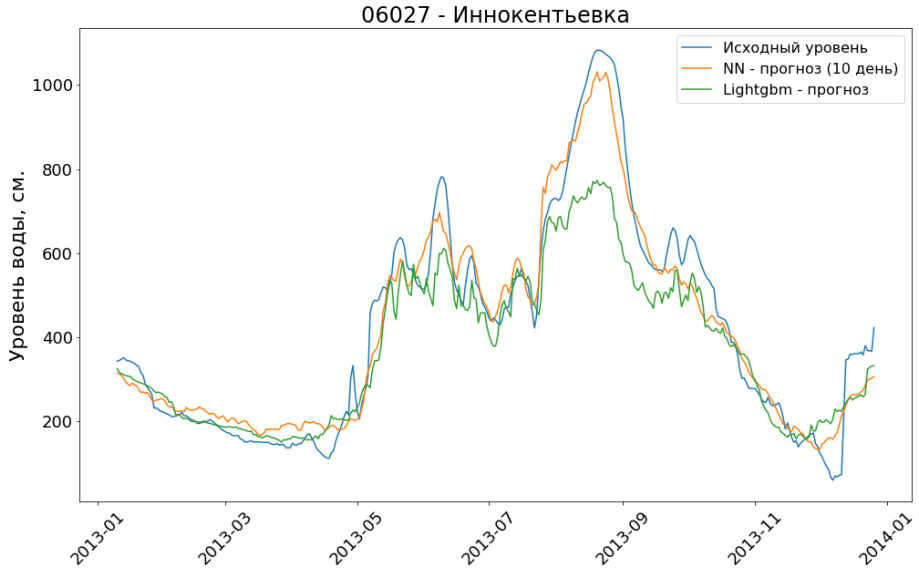


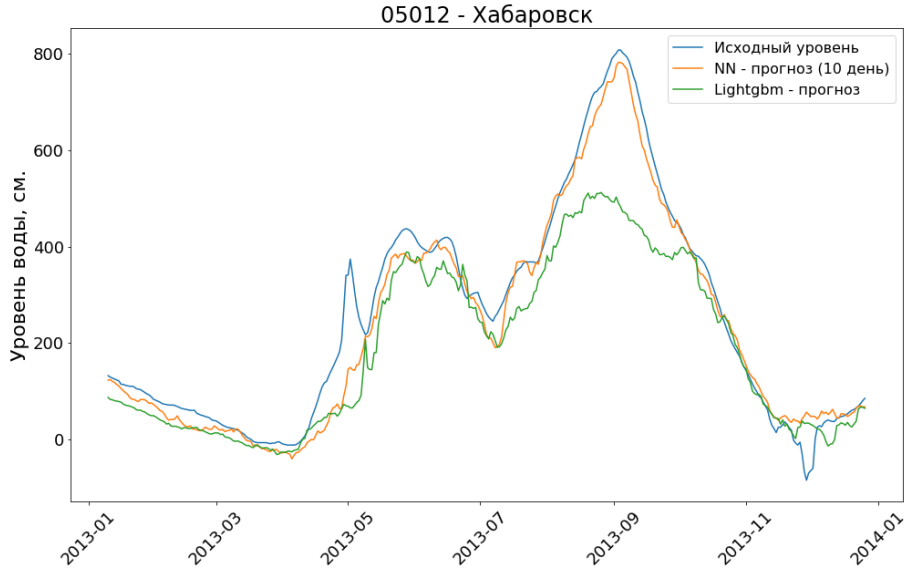


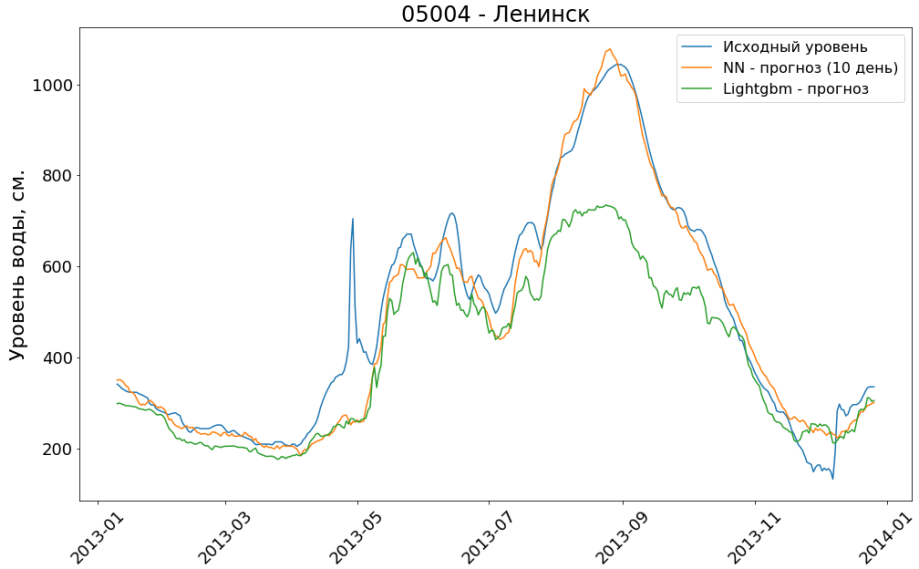
Сравнение Lightgbm и нейросети на примере 2013 г

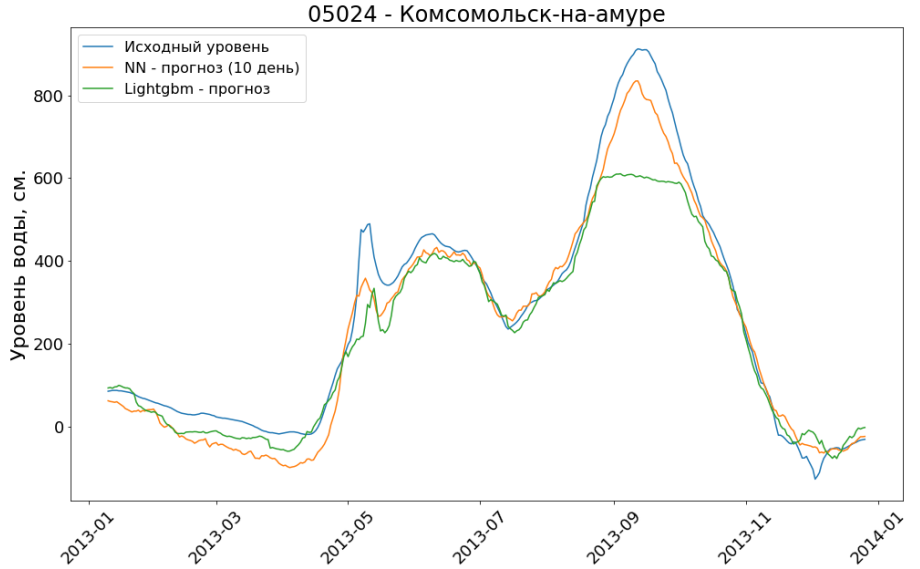
Обучение: все данные до 2013-01-01

Тест: 2013-01-01 – 2014-01-01









Результаты прогноза на 5 лет вперед

Обучение: все данные до 2013-01-01 **Тест:** 2013-01-01 – 2017-12-21

Преимущества использования такой модели

- Прогнозирование уровня сразу для множества станций – количество может легко регулироваться и учет взаимозависимости уровней между гидропостами. Так же за счет weight sharing одновременное прогнозирование на множестве гидропостов можно рассматривать как дополнительную регуляризацию.

- Простое обобщение на другие гидропосты и реки – алгоритм можно легко обобщить на бОльшее количество гидропостов, и без особных сложностей запустить на других крупных речных бассейнах

- Требуется меньший объем данных по сравнению с другими методами (линейные модели, деревья, случайный лес, бустинг) – можно брать историю по 10 дней с шагом в 1 день, увеличивая выборку (это показывает улучшение качества) – например, если в выборке всего история за месяц, то для такой модели мы можем это развернуть в 20х10 обучающих примеров – 20 примеров, в каждом из которых история по 10 дней

- Архитектура нейросети подобрана для этой задачи таким образом, чтобы признаки в нейросети были максимально приближены к реальным гидрофизическим и метеорологическим характеристикам и процессам (на основе изученной литературы)

- Модель хорошо обобщается на случаи опасных повышений уровня (на примере 2013 г.)

- Возможность подключения новых моделей по прогнозу погоды. Кроме пессимистического сценария можно одновременно рассматривать оптимистичный и нейтральный и следить за изменением уровня при разных вариантах развития событий.

Что еще можно улучшить в модели

- Категориальные переменные (облачность, погода между сроками, код состояния водного объекта) – эти данные тоже обладают полезной информацией, особенно код состояния водного объекта в зимний период с информацией о поведении льда, но на текущий момент эти данные исключены из рассмотрения (так как не вся информация присутствует в выборке, и неизвестно как это скажется во время тестирования)

- Подключить дополнительные метео-признаки (температура почвы, в АИСОРИ есть еще другие датасеты, не стал их включать в текущий датасет т.к. история в АИСОРИ ограничена и на момент тестирования вряд ли будут обновлены данные)

- Попробовать графовые нейросети (напр. http://snap.stanford.edu/graphsage/) – между гидропостами по реке можно составить граф связей и такие нейросети выглядят в этой задаче потенциально интересными.

- Подключить новые модели прогноза погоды (напр. Gismeteo, AccuWeather)

Список изученной литературы

- [1] Спектральный анализ временных рядов в экономике [Текст] / К. Гренджер, М. Хатанака ; Пер. В. С. Дуженко, Е. Г. Угер ; Науч. ред. В. В. Налимов. - Москва : Статистика, 1972. - 312 с. : черт.; 22 см.
- [2] С.В. Борщ, Ю.А. Симонов, А.В. Христофоров, Н.М. Юмина. КРАТКОСРОЧНОЕ ПРОГНОЗИРОВАНИЕ УРОВНЕЙ ВОДЫ НА РЕКЕ АМУР
- [3] Известия Томского политехнического университета. Инжиниринг георесурсов. 2016. Т. 327. № 11. 105–115 Лариошкин В.В. Методика прогноза дождевых паводков в бассейне Верхнего Амура (на примере р. Онон)
- [4] Экстремальные паводки в бассейне Амура: гидрологические аспекты / Сб. работ по гидрологии / под ред. Георгиевского В.Ю., ФГБУ «ГГИ», СПБ, ООО «ЭсПэХа», 2015.- стр.171.
- [5] Мы и амурские наводнения: невыученный урок? / Под ред.А. В. Шаликовского. — М.: Всемирный фонд дикой природы (WWF),2016. — 216 с.
- [6] Калугин А.С., Модель формирования стока реки Амур и ее применение для оценки возможных изменений водного режима, дис. … канд. геогр. Наук. Институт водных проблем РАН, Москва, 2016
- [7] С.В. Борщ , Д.А. Бураков , Ю.А. Симонов. МЕТОДИКА ОПЕРАТИВНОГО РАСЧЕТА И ПРОГНОЗА СУТОЧНОГО ПРИТОКА ВОДЫ В ВОДОХРАНИЛИЩЕ ЗЕЙСКОЙ ГЭС
- [8] ЭКСТРЕМАЛЬНОЕ НАВОДНЕНИЕ В БАССЕЙНЕ АМУРА В 2013 ГОДУ: АНАЛИЗ ФОРМИРОВАНИЯ, ОЦЕНКИ И РЕКОМЕНДАЦИИ, Болгов М.В., Алексеевский Н.И., Гарцман Б.И., Георгиевский В.Ю., Дугина И.О., Ким В.И., Махинов А.Н., Шалыгин А.Л. География и природные ресурсы. 2015. № 3. С. 17-26.