Прогнозирование рядов солнечной и геомагнитной активности рекуррентными нейросетями

Б. В. Козелов

Полярный геофизический институт

Апатиты, Россия

Аннотация

В моделях ионосферы и верхней атмосферы, важных для многих прикладных вопросов, используются индексы солнечной и геомагнитной активности, которые получаются на основе наблюдений наземными приборами и спутниками в солнечном ветре. Прогнозирование рядов геомагнитной активности вперед с некоторой точностью решается при наличии данных о солнечной активности и о солнечном ветре. Трудно формализуемые связи могут быть включены в модель с помощью нейросетевого подхода.

В имеющихся наборах данных о солнечном ветре (база OMNI) имеются пропущенные значения (~10%), которые создают проблемы для использования нейросетей. Простые стандартные методы заполнение отсутствующих значений, такие как использование медианы или среднего значения, нарушают статистические характеристики рядов и не всегда могут работать успешно.

В докладе обсуждаются следующие задачи:

1. Заполнение отсутствующих значений в рядах данных межпланетного поля (ММП) Btot, Bz, V, Np по информации об этих величинах за несколько дней и текущих индексов геомагнитной активности SYM-H, AL, AU, AE, Kp.
2. Прогнозирование рядов индексов солнечной активности (число солнечных пятен и F107) с использованием слоев LSTM.
3. Прогнозирование рядов индексов геомагнитной активности SYM-H, AL, AU, AE, Kp на основе данных о предыдущей активности, рядов индексов солнечной активности и параметров солнечного ветра.

Заполнение пропусков в рядах данных OMNI

Солнечные данные:

R – число пятен
F10.7 – поток радиоизлучения

Других постоянных источников информации нет

Заполнение интерполяцией по соседним значениям

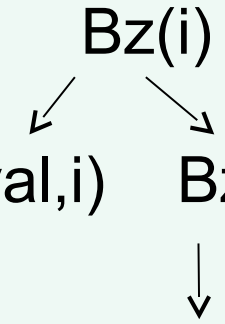
Параметры солнечного ветра:

Bz – z-компонента ММП
P – давление солнечного ветра
Np – плотность частиц
V — скорость солнечного ветра

Ряды геомагнитных индексов не имеют пропусков

Заполнение значениями, исходя из корреляций с AL, AU и SYM-H

Алгоритм заполнения для (5-минутного) ряда Bz



Вектор признаков в рядах AL, AU, SYM-H (где это значение Bx(i) может влиять):

$C\_vec(j) = AL(j+c_1, \ldots, j+d_1), AU(j+c_2, \ldots, j+d_2), SYM\text{-}H(j+c_3, \ldots, j+d_3)$,

куски рядов берутся из корреляций с Bz.

Ищем ближайший вектор Vec(i), min|C_vec(j)-Vec(i)| ,
для которого есть соответствующее значение Bz(val,i),
заполняем пробел j.

Временные масштабы корреляций для построения векторов признаков

	Bz	B	P	Np
AL	30-60 мин	30-120 мин	2-22 ч	16.5-24.5 ч
AU	30-60 мин	60-150 мин	0-10 ч	4-12 ч
SYM-H	30-120 мин	6.5-25 ч	10-32 ч	25-70 ч

Обсуждение

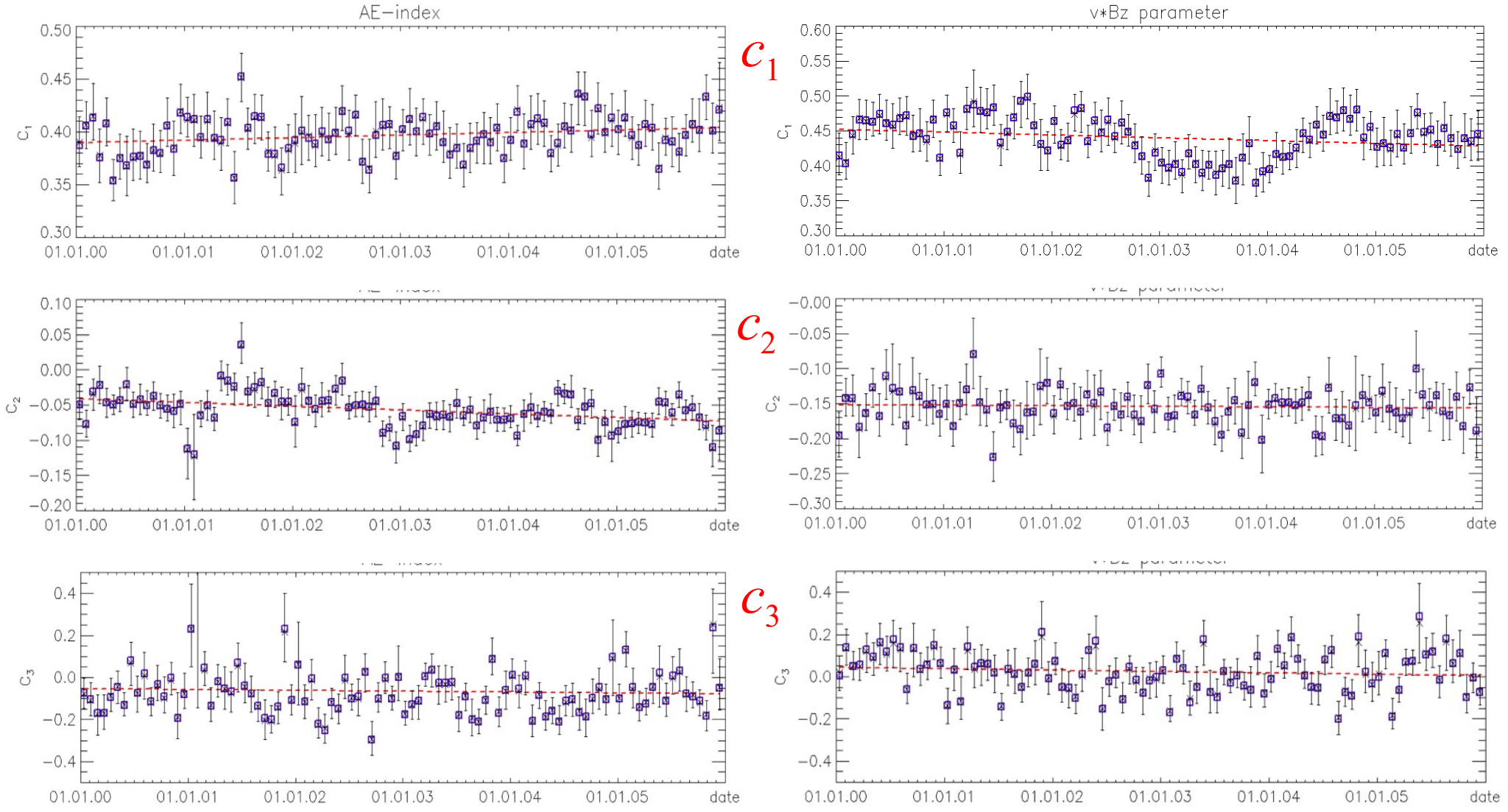
Доклад ранее – Kozelov B.«Wavelet leaders and bootstrap techniques for multifractal analysis», 2009.

Использовались куски ряда перед и после пробела, длинной в пробел *dl*. В ряду искался кусок длиной 3*dl*, который начинался и заканчивался так же (наиболее близко).

Сохранялись статистические свойства (мультифрактальный спектр)

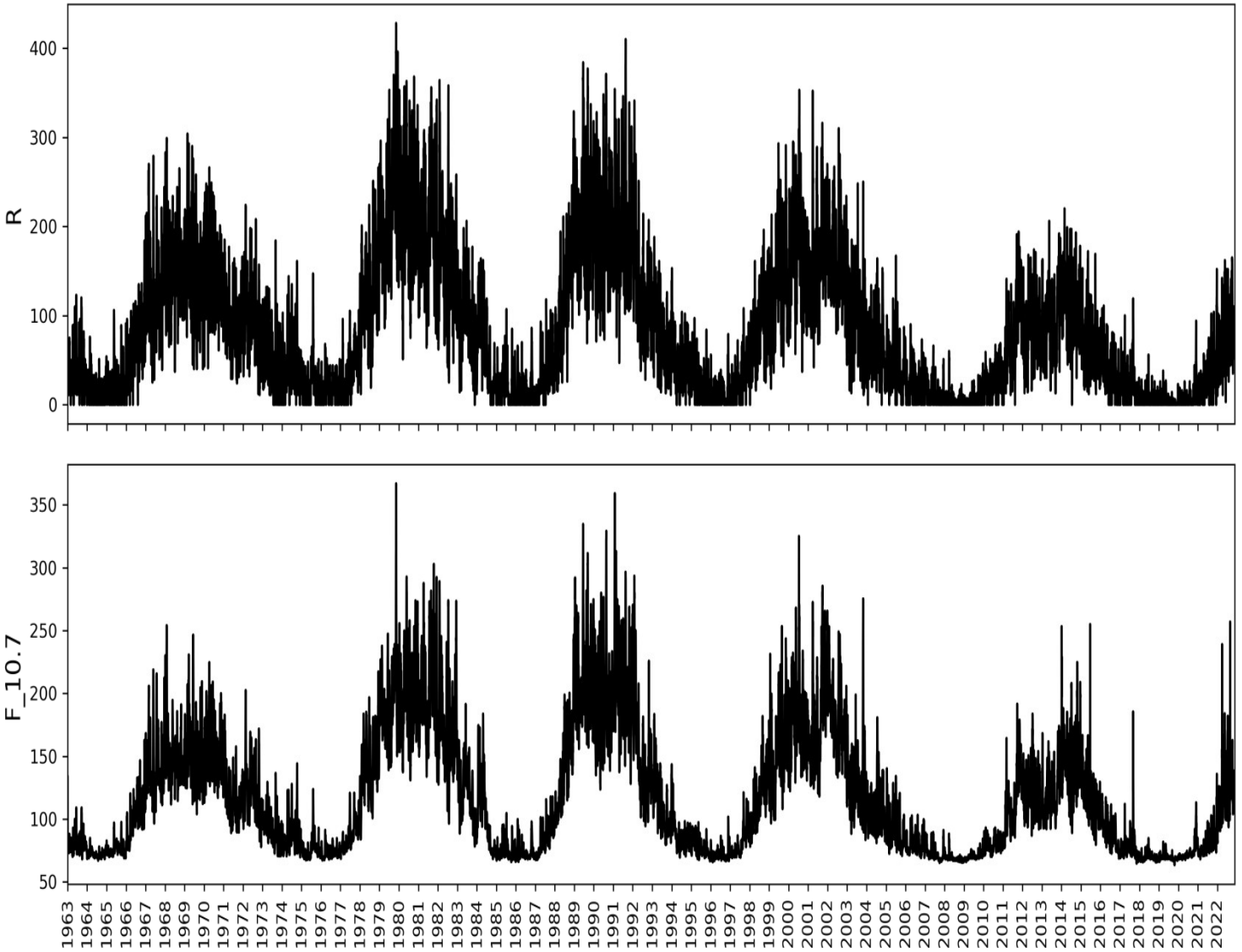
Недостаток: значения не согласованы со значениями геомагнитных индексов.

AE vB_z

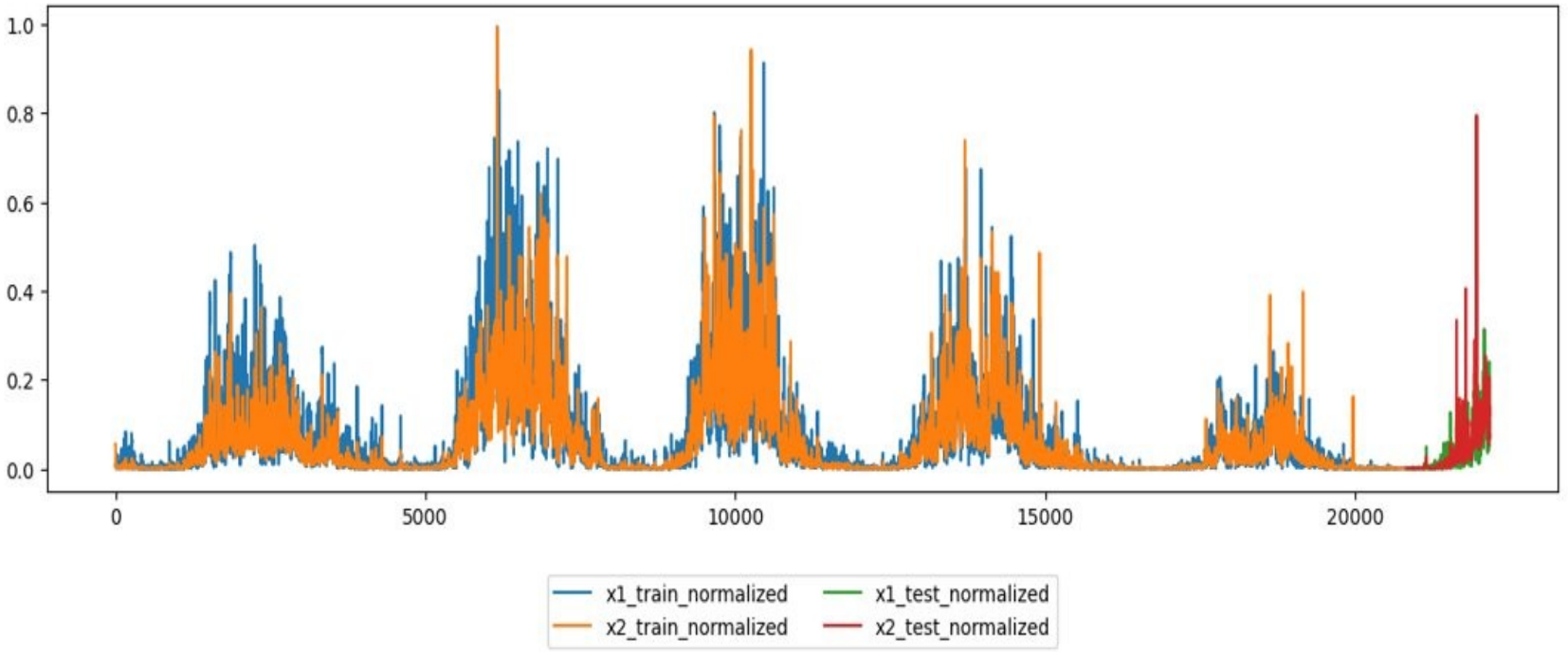

AE−index
v*Bz parameter
c_1
c_2
c_3
c_1
c_2
c_3
0.50
0.45
0.40
0.35
0.30
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
0.4
0.2
0.0
−0.2
−0.4
0.4
0.2
0.0
−0.2
−0.4
01.01.00
01.01.01
01.01.02
01.01.03
01.01.04
01.01.05
date
2000 2001 2002 2003 2004 2005
2000 2001 2002 2003 2004 2005

Индексы солнечной активности

Набор данных представляет собой суточные значения параметров, характеризующих солнечную активность, число солнечных пятен R и среднесуточное значение потока радиоизлучения на волне 10.7 см, который измеряется в солнечных единицах потока: 1 с.е.п. = 10^{-22} Вт/(м^2 Гц).
Временные ряды с 1 января 1961 года по 1 марта 2023 года взяты из базы данных OMNI.
Известно, что оба ряда коррелируют между собой и содержат солнечные периодичности: 11 летний солнечный цикл и ~28 дневное собственное солнечное вращение. Отсутствующие отдельные значения во временном ряду были интерполированы по соседним.



Нормировка данных



LSTM — нейронная сеть с долгой краткосрочной памятью

Люди не запускают мыслительный процесс с нуля в каждый момент времени. Читая статью, вы понимаете смысл каждого слова на основе значений предыдущих слов. Мысли имеют свойство накапливаться и влиять друг на друга. Этот принцип используется в сетях рекуррентных нейронных сетях (РНС).

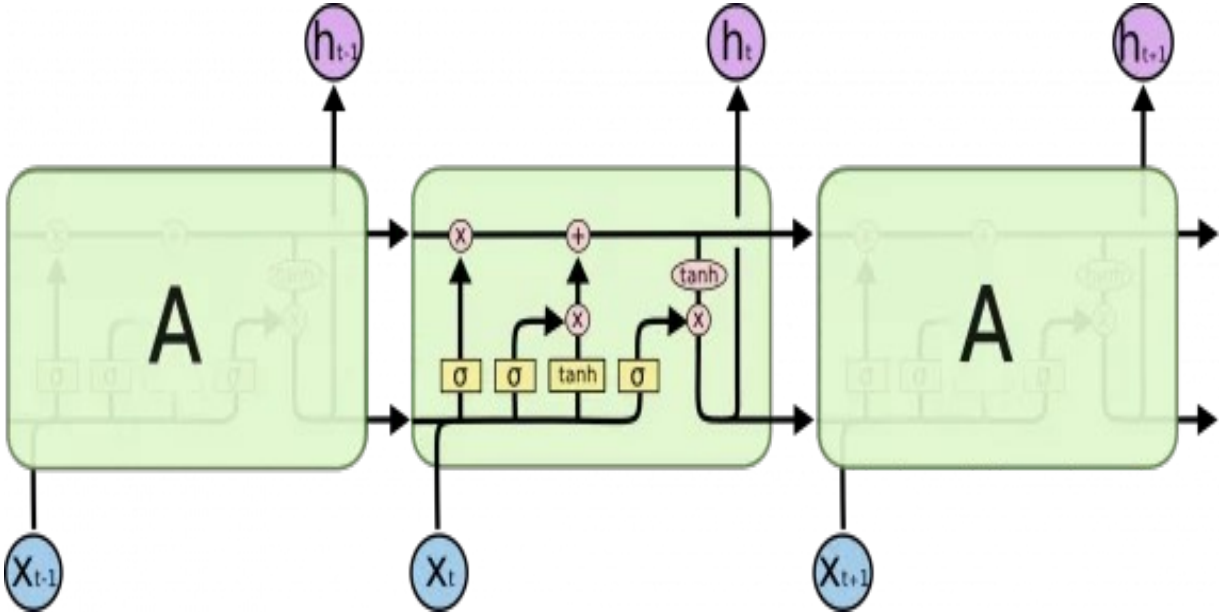
В теории РНС способны справиться с такими «долгосрочными зависимостями». К сожалению, на практике РНС не способны решить эту задачу. Проблема подробно исследована в работах [Hochreiter (1991)] и [Bengio, et al. (1994)], в которых выявлены фундаментальные ограничения РНС.

Существенным продвижением стали LSTM — специфического типа рекуррентные нейронные сети, которые решают отдельные задачи гораздо эффективнее стандартных методов.

LSTM (long short-term memory, дословно (долгая краткосрочная память) — тип рекуррентной нейронной сети, способный обучаться долгосрочным зависимостям. LSTM были представлены в работе [Hochreiter & Schmidhuber (1997)], впоследствии усовершенствованы другими исследователями

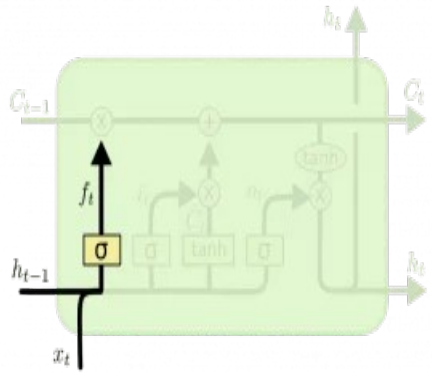
В работе использована реализация слоя LSTM в библиотеке keras в пакете TensorFlow

LSTM — нейронная сеть с долгой краткосрочной памятью

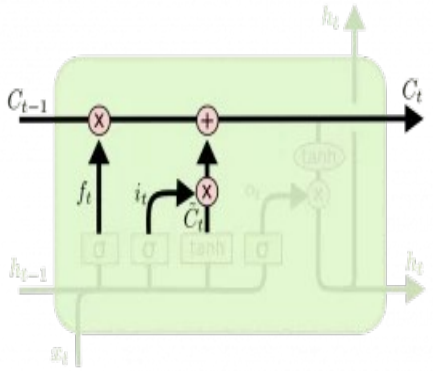


Повторяющийся модуль LSTM состоит из четырех взаимодействующих слоев

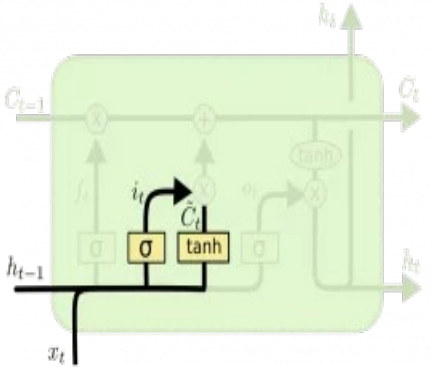
LSTM — нейронная сеть с долгой краткосрочной памятью



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

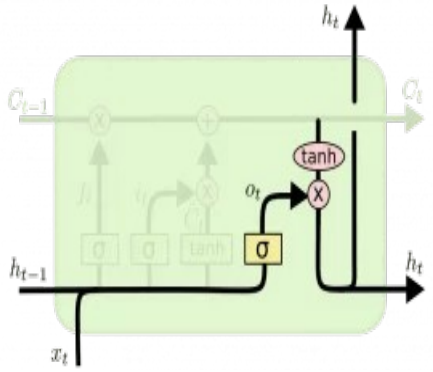


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$



$$o_t = \sigma\left(W_o \left[h_{t-1}, x_t\right] \; + \; b_o\right)$$

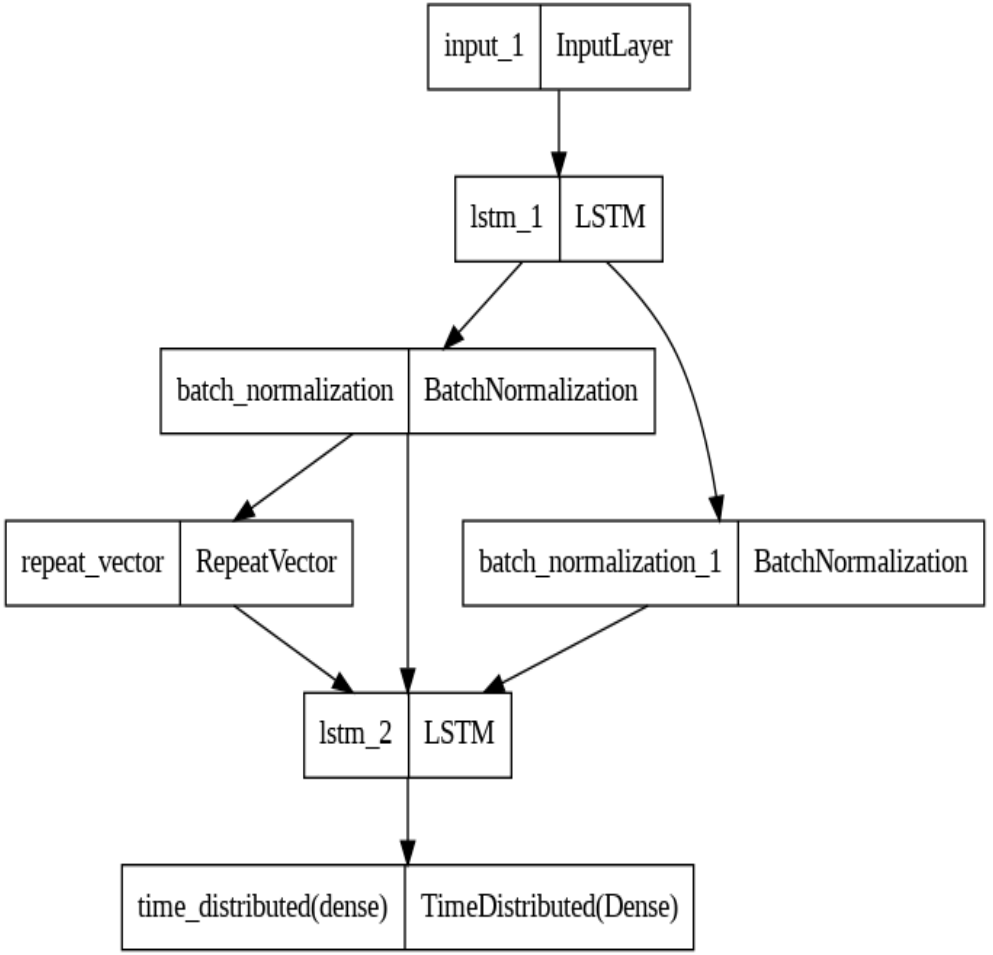
$$h_t = o_t * \tanh\left(C_t\right)$$

Нейронная сеть для предсказания значений по предыстории

Для построения численной модели использована реализация LSTM (Long short-term memory) слоя на языке Python в пакете keras библиотеки tensorflow.

Сеть обучалась по данным за 5 солнечных оборотов (140 земных дней) предсказывать значения вперед на 28 дней. Для этого из всего ряда данных сформированы массивы векторов из 140 значений для подачи на вход сети, и соответствующие им массивы векторов из 28 истинных следующих за ними значений для подачи на выход сети во время тренировки.

Подаваемые на вход значения последовательно кодируются во внутреннем состоянии первой LSTM сети, которое подается на вторую LSTM сеть, которая предсказывает следующие значения. Эти значения сравнивались с истинными, в процессе обучения минимализировалась метрика «средняя абсолютная ошибка» MAE – mean absolute error). В данном случае использовались две LSTM-сети по 140 нейронов.

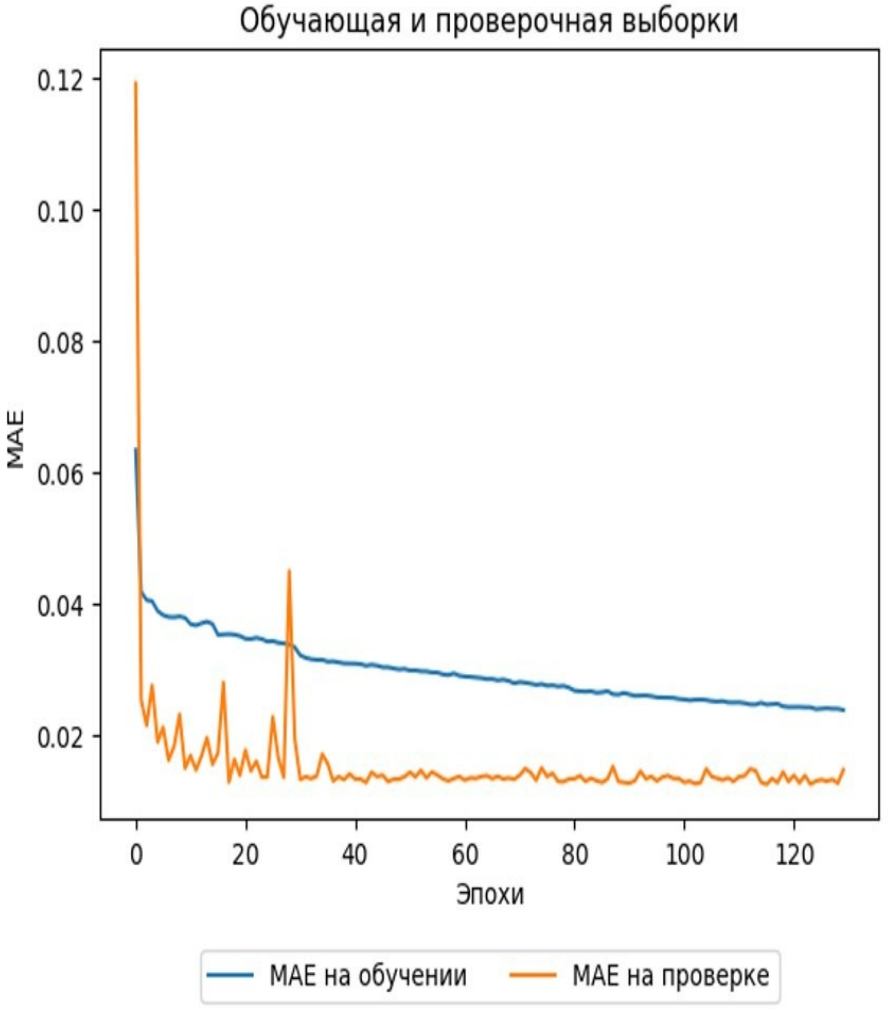


Обучение

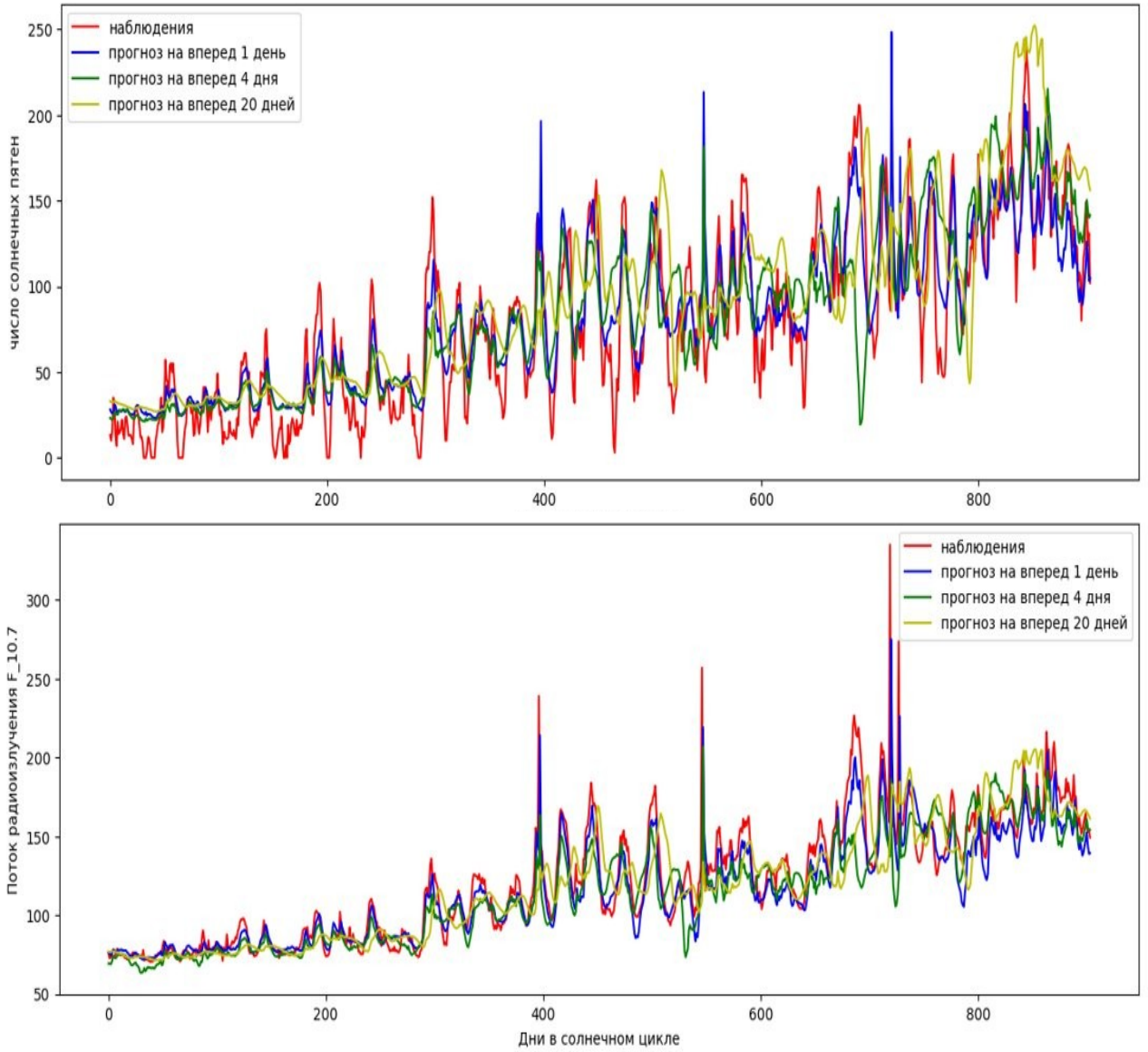
Сеть обучалась по 140 значениям предсказывать 28 следующих.

130 эпох

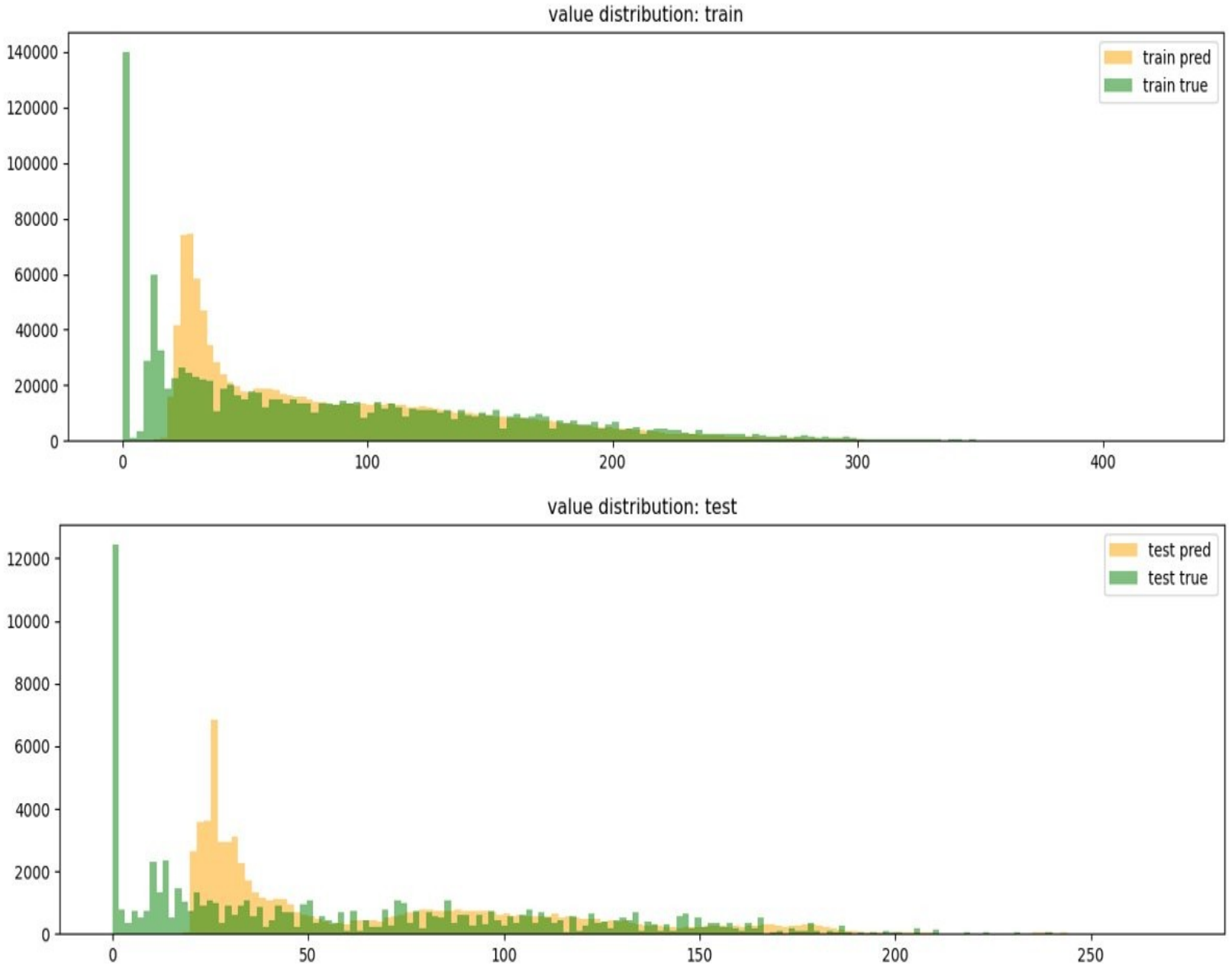
Средняя абсолютная ошибка предсказания модели составляет менее 2 %.



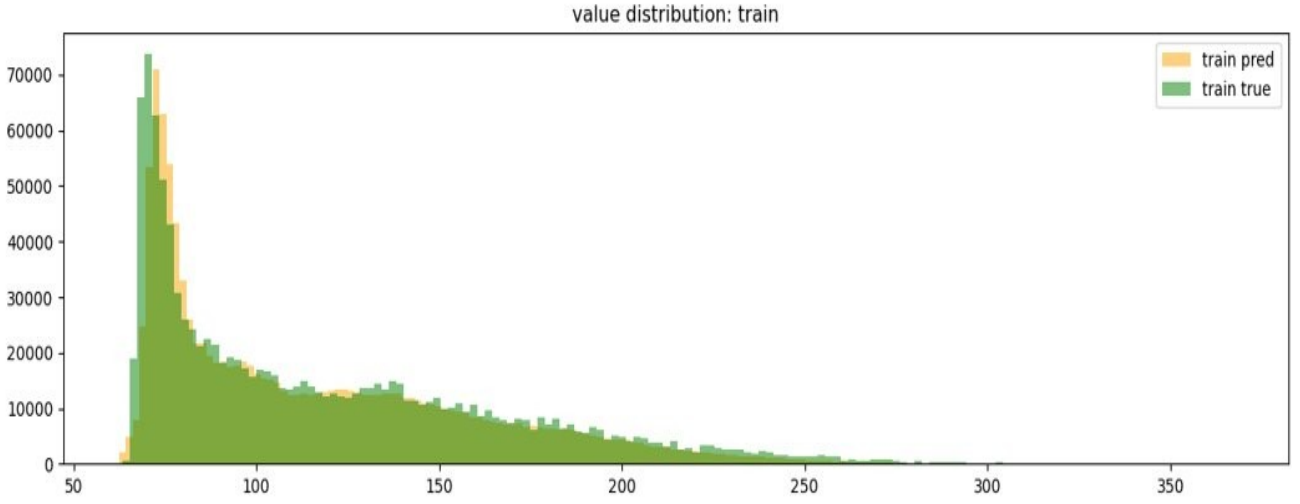
Сравнение на тестовых данных

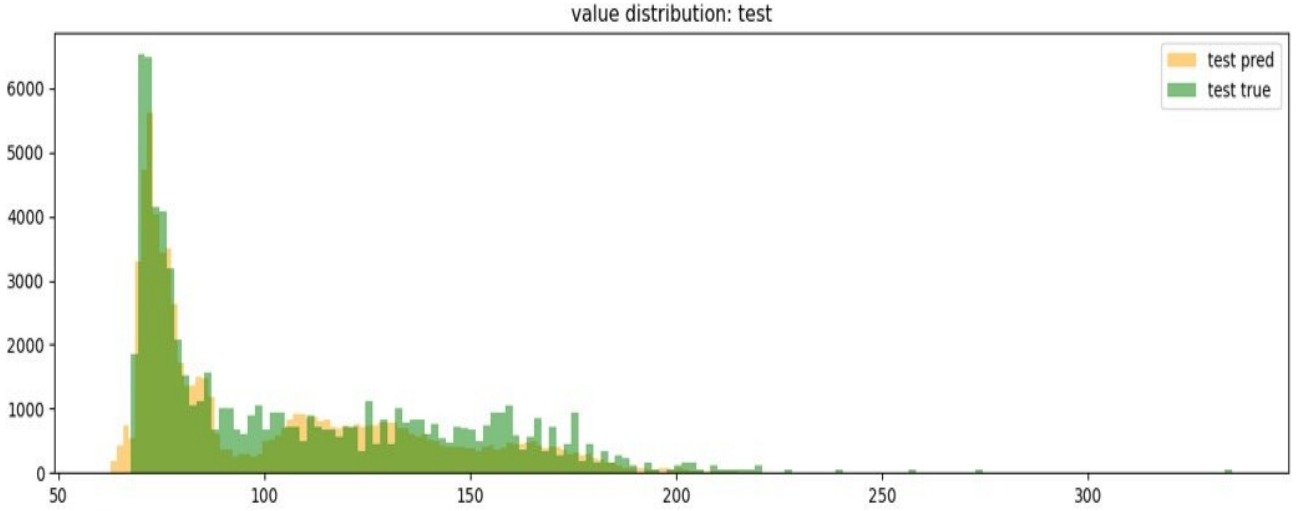


Сравнение распределений для R



Сравнение распределений для $F_{10.7}$





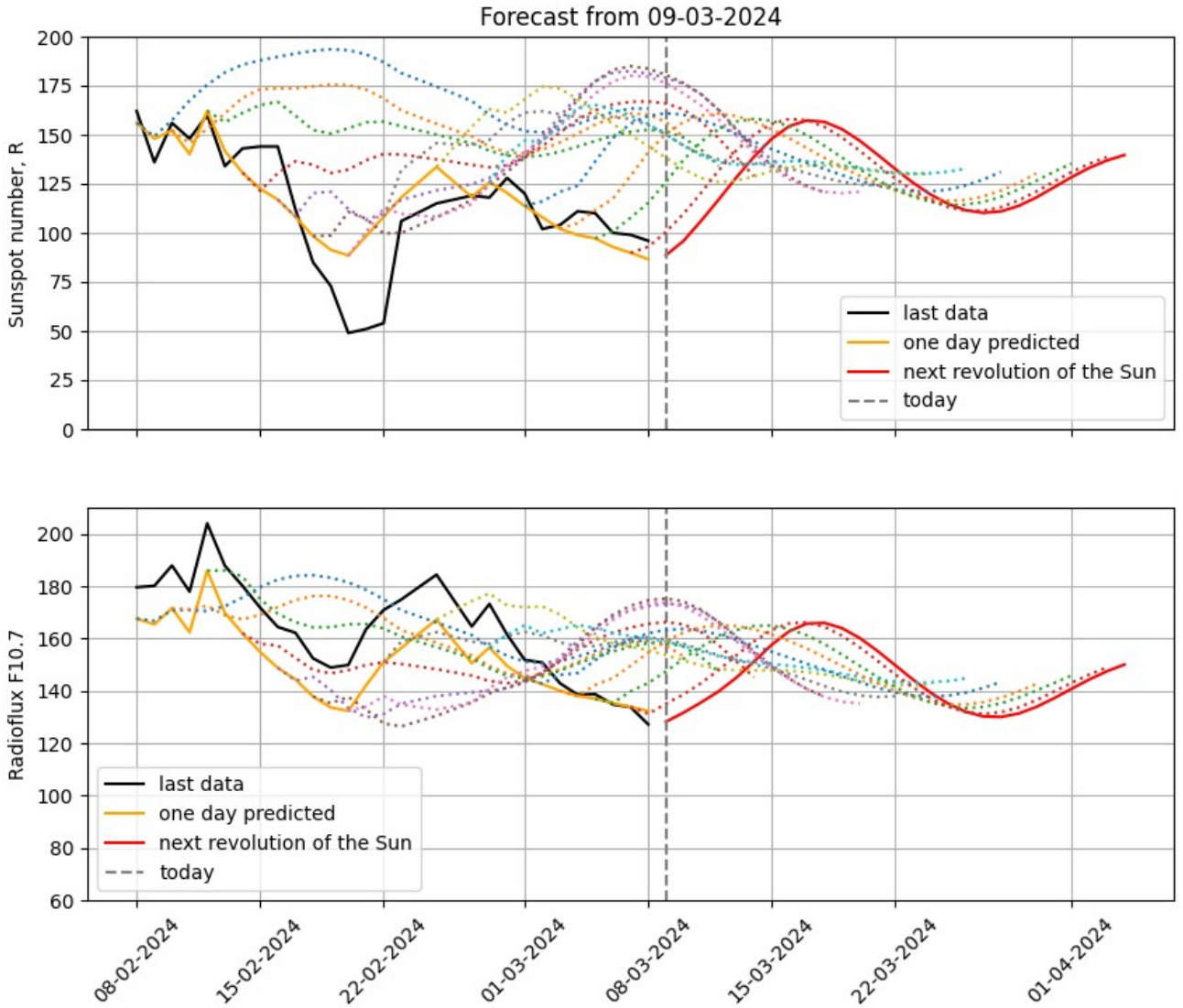
Реализация на сайте

Ссылка:

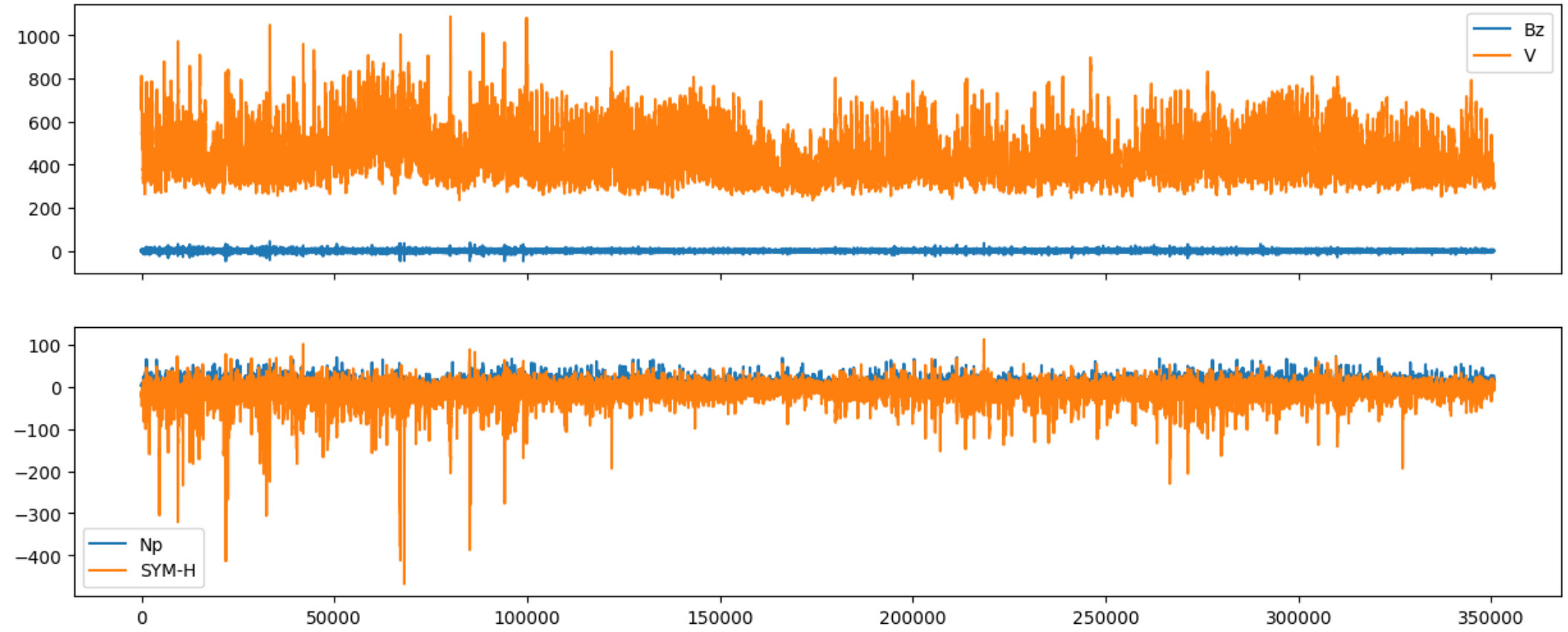
http://aurora.pgia.ru/AI/?id=R_f107

Прогноз обновляется раз в сутки.

Текущие значения загружаются с web-сайтов



OMNI dataset 2000-2019



Nonmalisation:

Bz	V	Np	SYM-H
44.25,	1086.6,	69.41,	112.0
-48.01,	234.9,	0.04,	-468.0

Arrays:

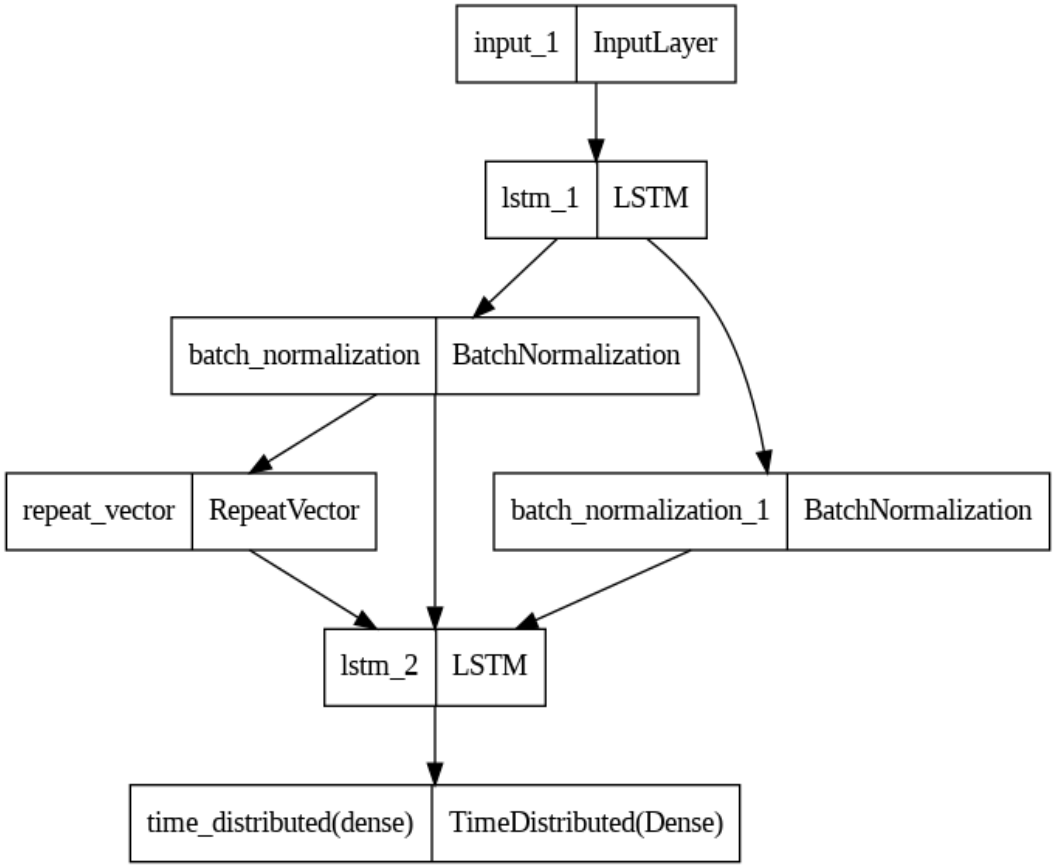
Train	(297844, 200, 5)	(297844, 24, 2)
Val	(52573, 200, 5)	(52573, 24, 2)

Нейронная сеть для предсказания значений SYM-H по предыстории значений [Bz, V, Np, SYM-H]

Для построения численной модели использована реализация LSTM (Long short-term memory) слоя на языке Python в пакете keras библиотеки tensorflow.

Сеть обучалась на данных OMNI 2000 - 2019 гг. (из ряда 5-минутных значений с заполненными пробелами взяты значения через 30 минут) по предыстории [Bz, V, Np, SYM-H] за 100 часов предсказывать значения на 12 часов вперед.

Подаваемые на вход значения последовательно кодируются во внутреннем состоянии первой LSTM сети, которое подается на вторую LSTM сеть, которая предсказывает следующие значения. Эти значения сравнивались с истинными, в процессе обучения минимализировалась метрика «средняя абсолютная ошибка» MAE – mean absolute error). В данном случае использовались две LSTM-сети по 160 нейронов.

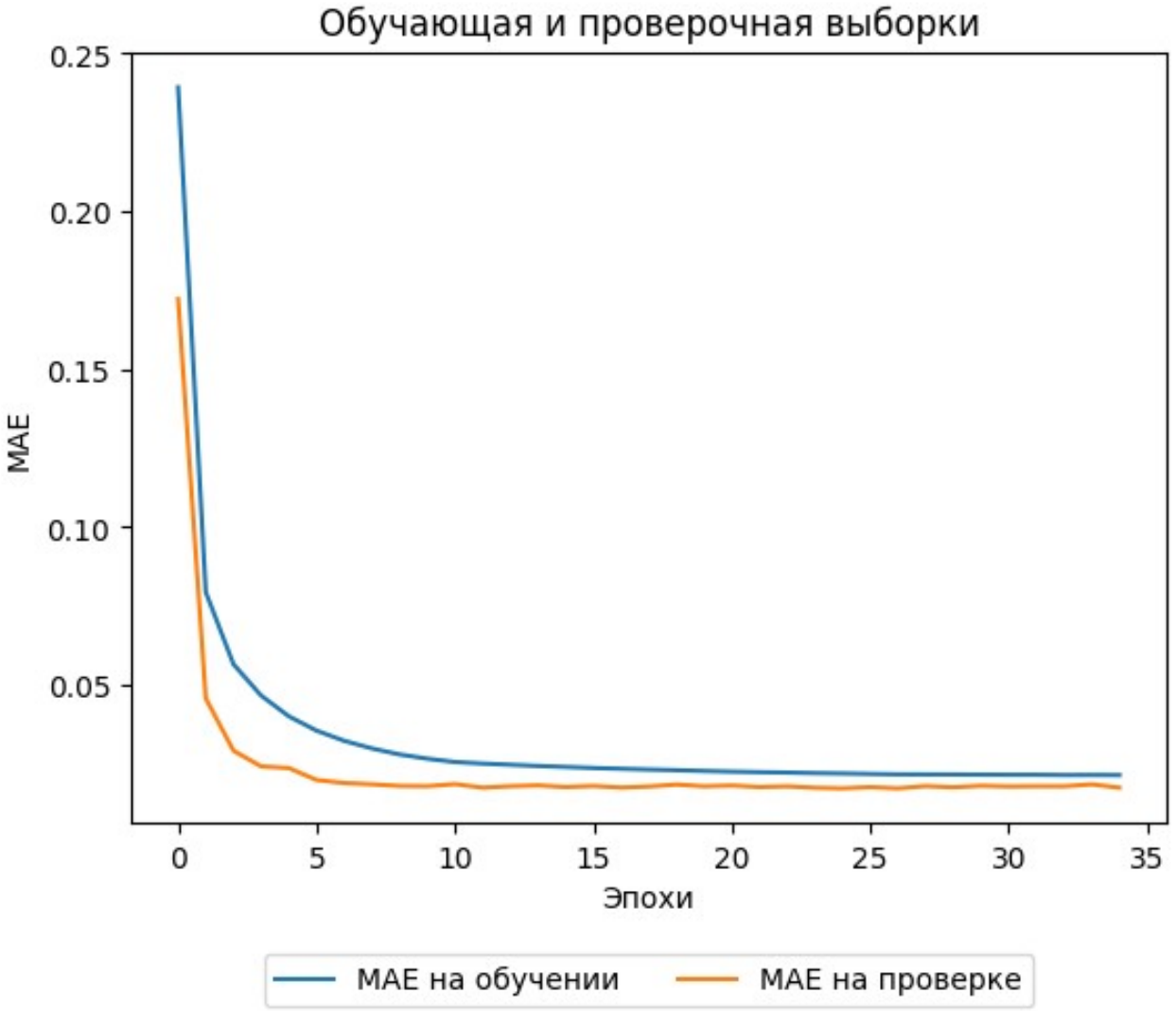


Обучение

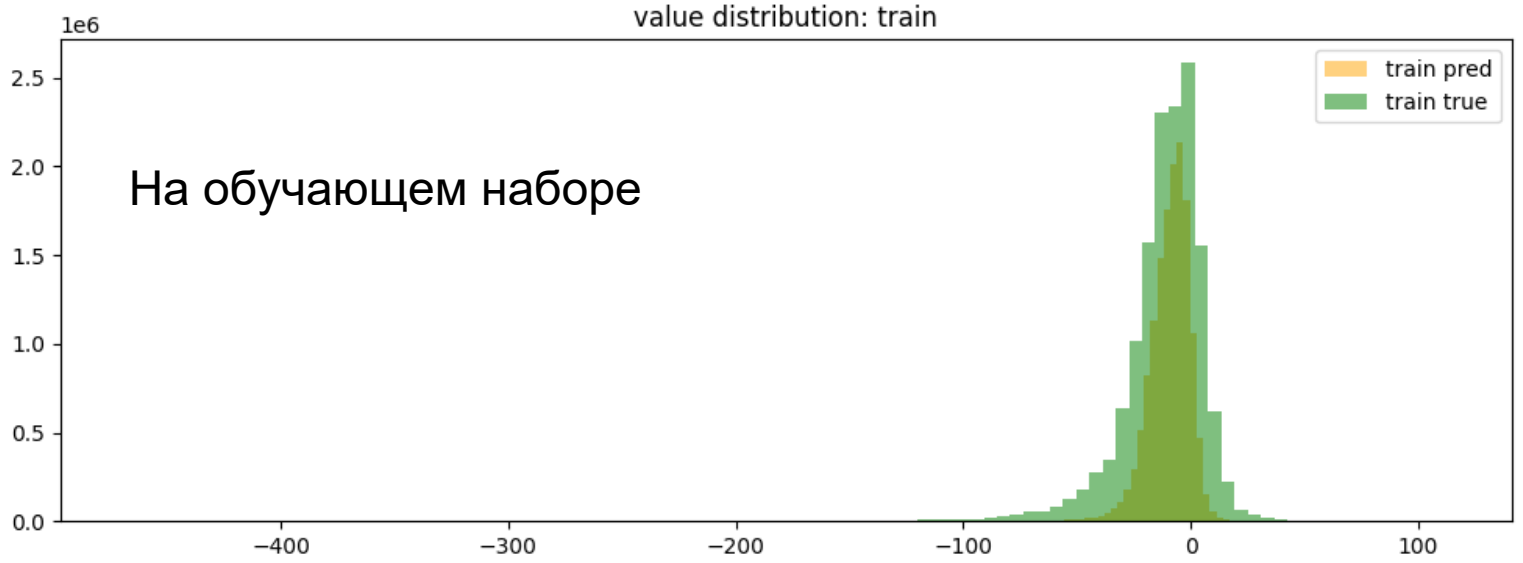
n_hidden = 160

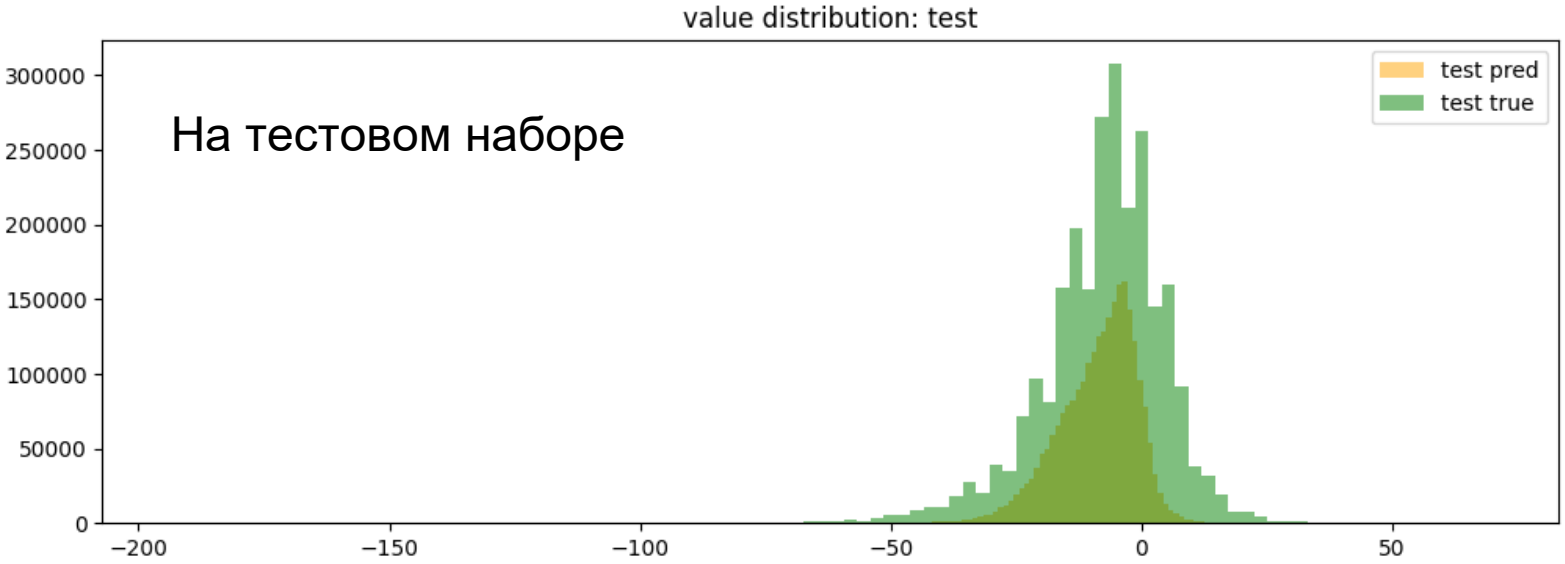
learning_rate уменьшалась от 1.e-5 до 1.e-6 при обучении

Средняя абсолютная ошибка предсказания модели составляет менее 2 %.

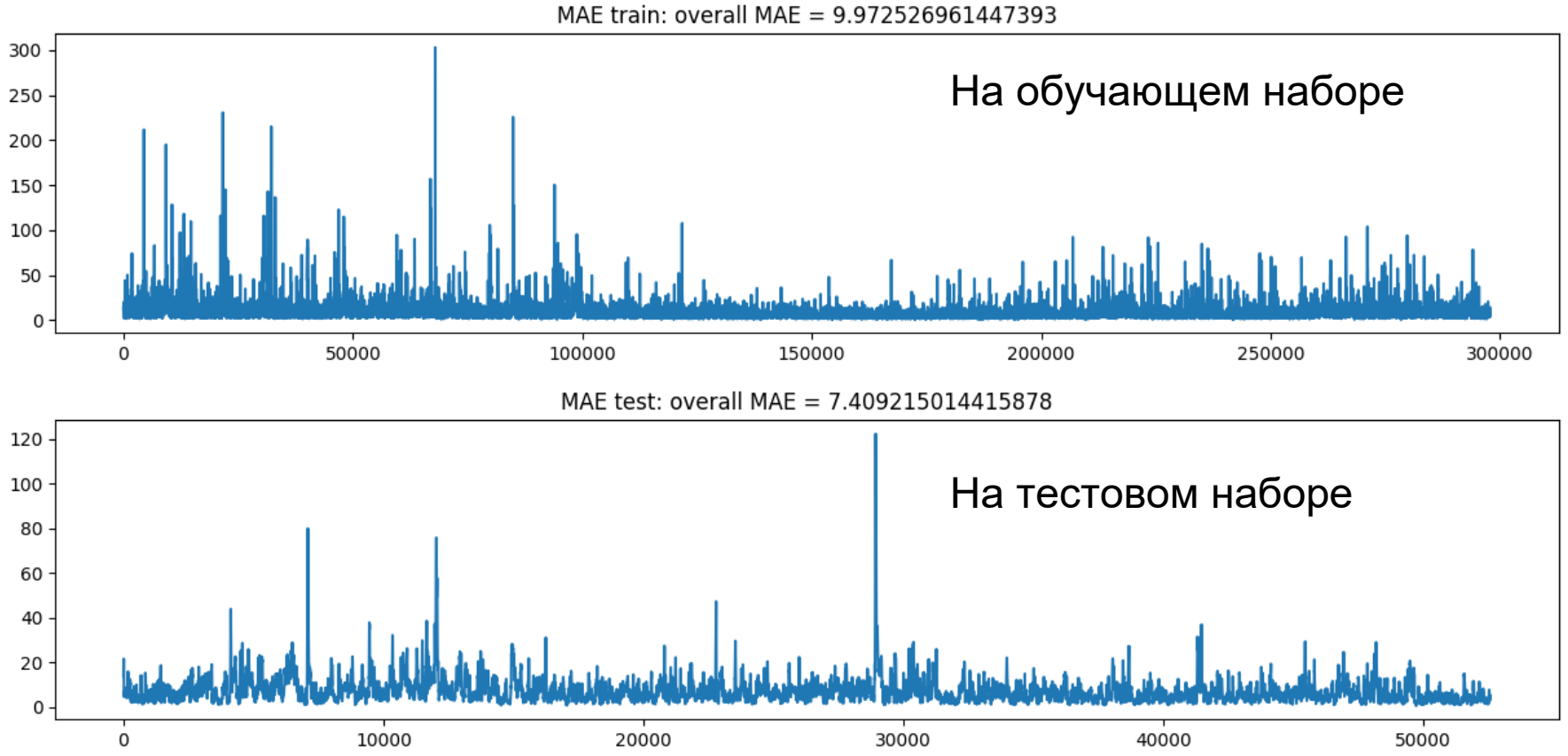


Сравнение распределений предсказанных и истинных значений

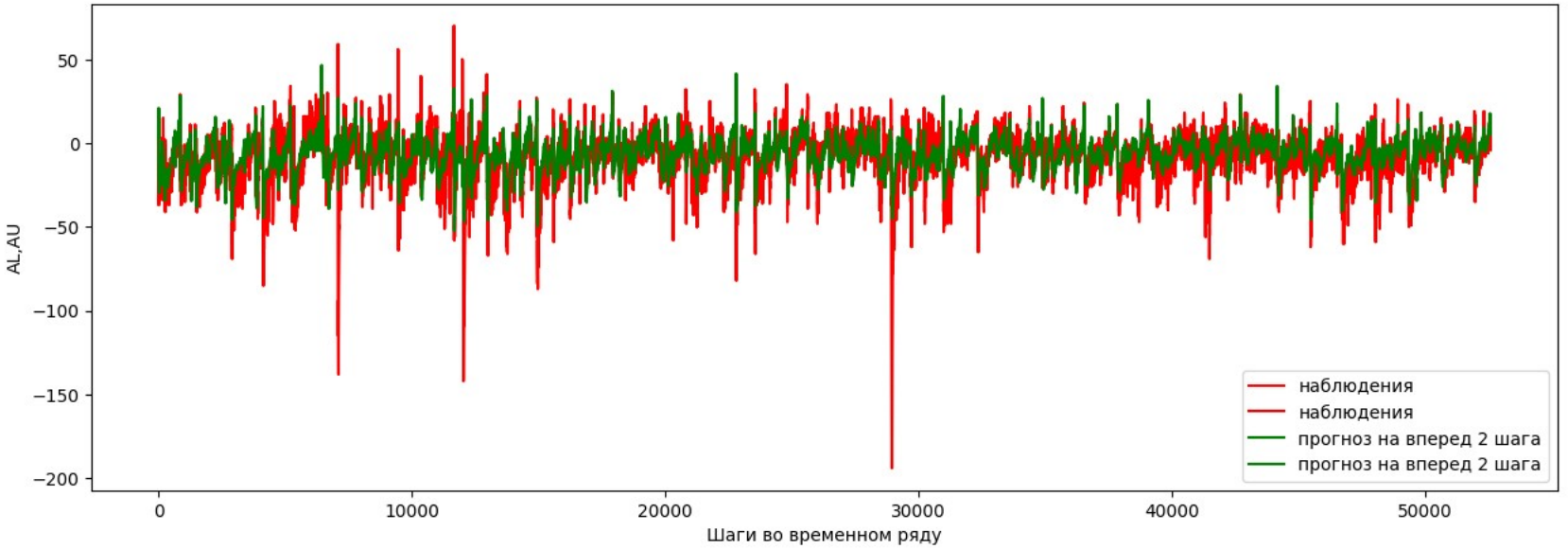


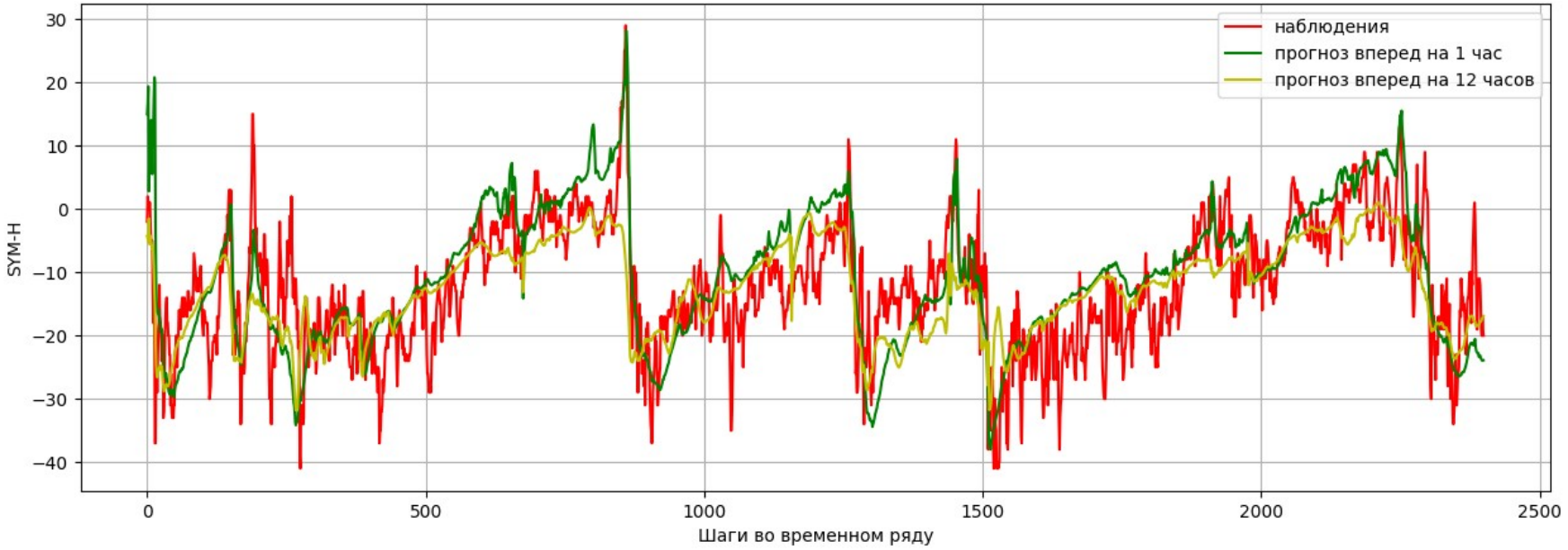


Абсолютная ошибка предсказания SYM-H в нТ

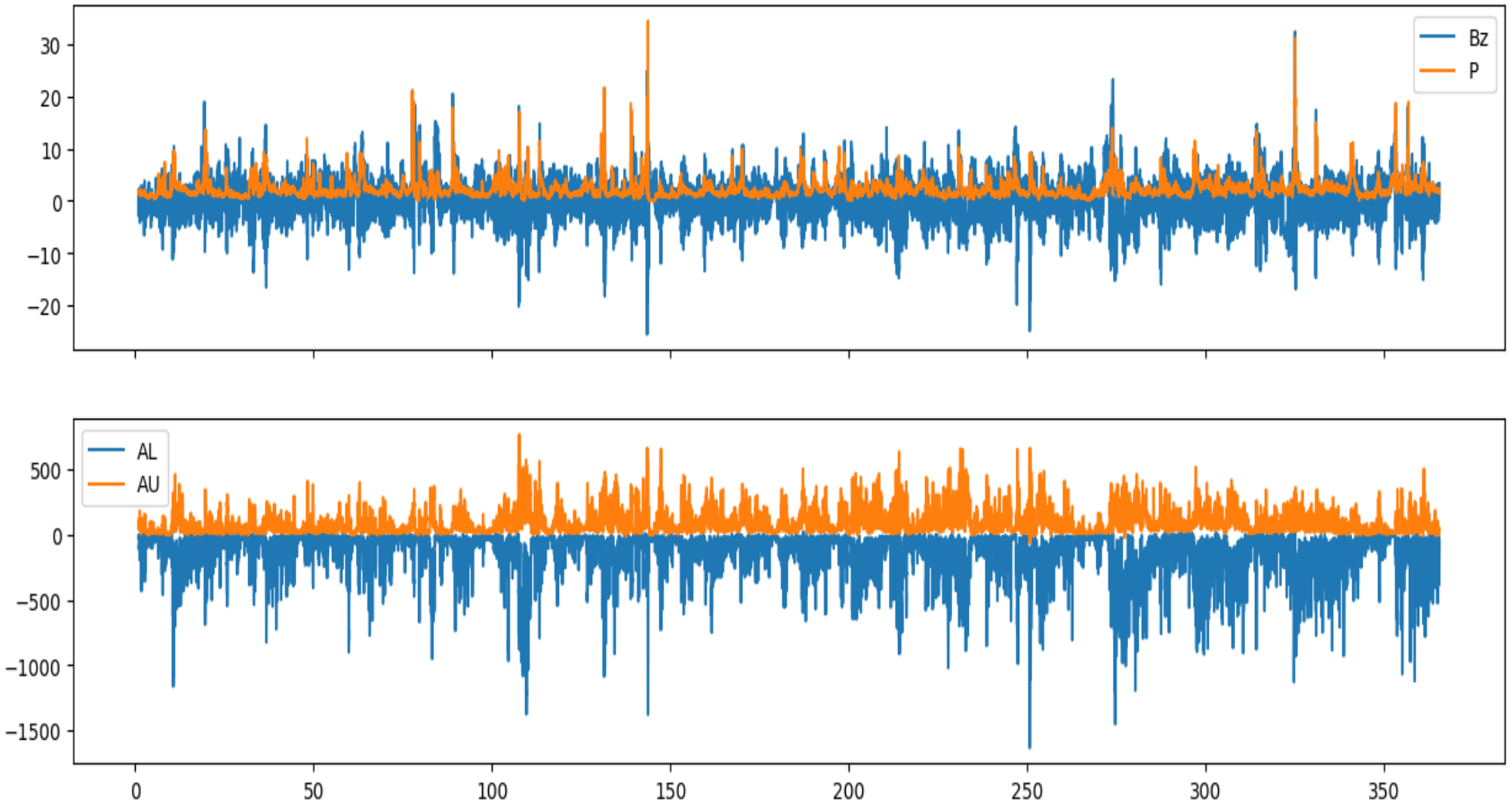


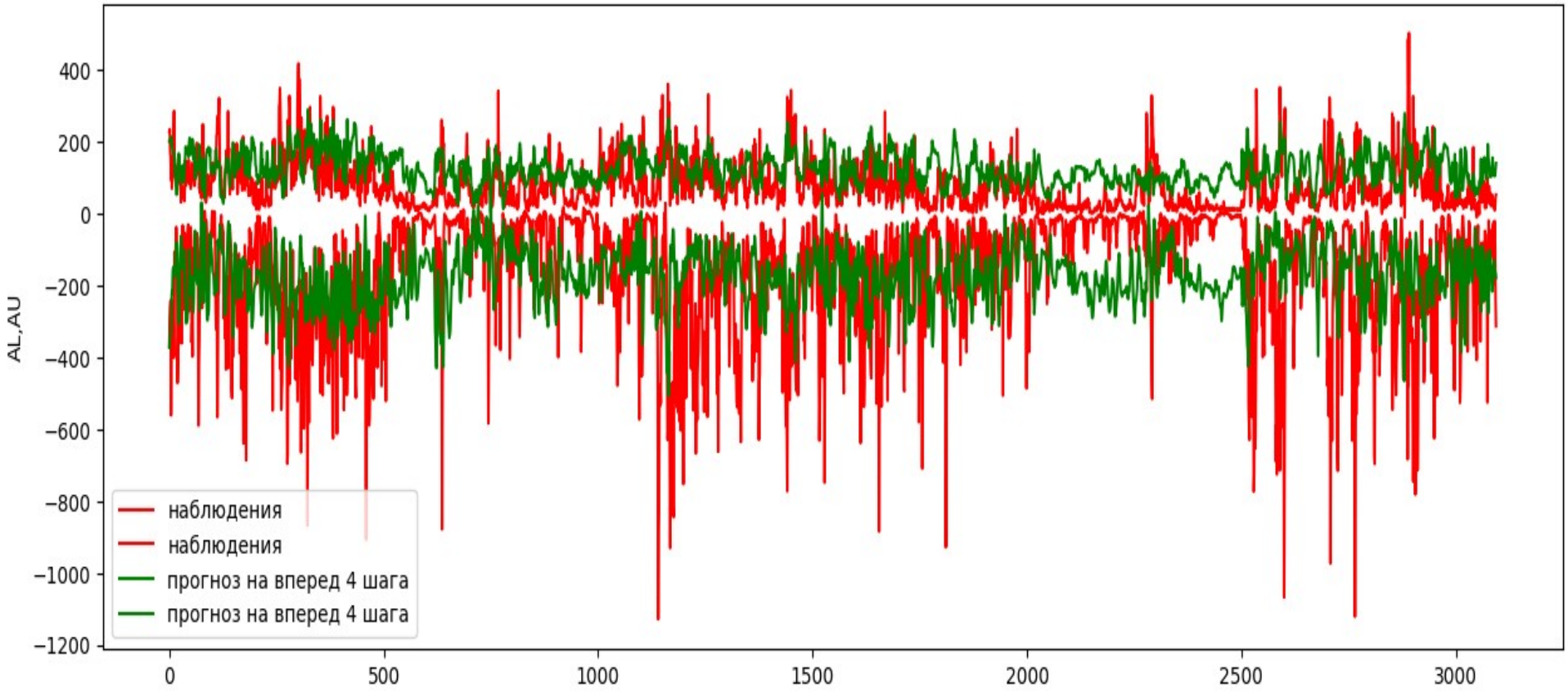
Примеры прогноза на тестовом наборе





Bz, P, AL, AU ⇨ forward AL,AU




AL,AU
400
200
0
−200
−400
−600
−800
−1000
−1200
0
500
1000
1500
2000
2500
3000
наблюдения
наблюдения
прогноз на вперед 4 шага
прогноз на вперед 4 шага

Выводы

1. Предложен алгоритм заполнения пропусков в рядах данных OMNI
2. Построена численная модель предсказания параметров солнечной активности — числа солнечных пятен R и потока радиоизлучения на волне 10.7 см F10.7 вперед на 28 суток. Численная модель использует искусственную нейронную сеть (ИНС) с LSTM (Long short-term memory) слоями. Средняя абсолютная ошибка предсказания модели составляет менее 2 %. Модель в реальном времени реализована на сайте http://aurora.pgia.ru и может быть дополнением к долгосрочным прогнозам других ИНТЕРНЕТ-ресурсов.
3. Построена численная модель предсказания значений на 12 часов. Численная модель использует искусственную нейронную сеть (ИНС) с LSTM (Long short-term memory) слоями. Средняя абсолютная ошибка предсказания модели составляет менее 2 %.

Работа поддержана РНФ, проект № 22-12-20017. Автор благодарит GSFC/SPDF OMNIWeb за подготовку использованных данных.