6th International Young Scientists Conference in HPC and Simulation, YSC 2017,
1-3 November 2017, Kotka, Finland

Absolute humidity anomalies and the influenza onsets in Russia: a computational study

Nikita E. Seleznev[a], Vasiliy N. Leonenko[a,*]

[a]ITMO University, 49 Kronverksky Pr. Saint-Petersburg, 197101, Russia

Abstract

In the current work we use a computational approach to analyze the association between the anomalous drops of absolute humidity and the subsequent onset of influenza epidemics in Russia. The correlation between these two factors is searched relying on the data of acute respiratory infection incidence in Saint Petersburg, Moscow and Novosibirsk. The analysis results, along with the output from the same analysis for Ile-de-France region (Paris and its suburbs), were compared with those achieved by Dr. Jeffrey Shaman and his co-authors for the US states. We show that although the analysis results for Ile-de-France are in agreement with the conclusions of Dr. Shaman, it does not hold true for the case of Russia, where the relation between the low levels of absolute humidity and the influenza onset is proved to be statistically insignificant.

1. Introduction

Influenza (or, shortly, flu) is one of the most common human infectious diseases. Being the most virulent among more than two hundred known seasonal acute respiratory infections (ARIs), it causes repetitive outbreaks resulting in high worker/school absenteeism and productivity losses. Also, if not treated properly, influenza leads to complications that may cause death of infected persons (the annual mortality rate is from 250 to 500 thousand individuals, according to WHO [12]).

To reduce the number of untreated cases of influenza it is crucial for the healthcare organs to be ready for the incoming epidemic. Although the seasonality of influenza outbreaks is acknowledged for a long time, its mechanism still does not have a satisfactory explanation. A lot of factors are named that may define the moment of an outbreak onset and its dynamics over time, but the extent of influence of one or another factor on the outbreak parameters is still arguable. Nevertheless, most of the researchers have come to terms with a fact that absolute humidity (AH)

* Corresponding author.
  E-mail address: vnleonenko@yandex.ru

dynamics may be correlated with epidemic occurrence [10], [11]. This statement was supported, among the other works, by results of data analysis for flu outbreaks in Russia [3], [6]. The experiment performed by Shaman et al. demonstrated a hands-on biological justification showing that virus survival and transmission among guinea pigs increased monotonically with a decrease in AH [8]. In the subsequent paper the authors extend their previous findings to the human population level, showing that the onset of increased wintertime influenza-related mortality in the US is preceded by anomalously low absolute humidity levels [9]. As a result, the authors state that "AH is a major (and likely predominant) determinant of influenza seasonality". At the same time, they admitted that some of their experiment results contradicted this statement: particularly, the association between the AH drop and a subsequent influenza onset did not reach statistical significance in a big part of the Northwestern USA.

In this paper we aim to answer the question if the anomalous drops in AH may precede the influenza onsets in Russian Federation using the long-term ARI weekly incidence data in three cities (Moscow, Saint Petersburg and Novosibirsk) from 1986 to 2015. For this sake we adapt the idea of Shaman et al. to use it for our data, implement the corresponding computational algorithm and perform the data analysis. The experiment results contribute to understanding whether the anomalous drops in absolute humidity before the flu outbreaks is inherent in case of Russia as well as in the USA and, thus, whether this phenomenon could be used to increase the quality of by far not-so-accurate model predictions obtained by the authors for flu dynamics in Russia [4], [5]. To perform the data analysis we have reconstructed the algorithm which aims at representing the average AH dynamics before and after influenza onsets and assesses the statistical significance of potentially abnormal AH dynamics.

2. Algorithm structure

2.1. Assessing AH dynamics before and after an onset

First of all, we need to form a dataset of *AH* deviations before and after an onset. This is done according to the following sequence of operations:

- Form an array which contains all the days of influenza onsets corresponding to the selected year period and geographical locations (states, cities, countries) under consideration. These days may be taken from external sources or calculated by the algorithm itself based on a fixed influenza outbreak criterion. Let *n* be the overall number of onset days.
- For $i \in \overline{1, n}$:
    - Take a sample of *AH* values corresponding to continuous measurements during 6 weeks (42 days) before and 4 weeks (28 days) after the *i*-th onset for the corresponding year and a state (70 values in total).
    - Using the *AH* sample, form a dataset $Z^{(i)}$ which contains corresponding values of AH', local daily deviations of AH from its 31-year mean for each day (70 values in total).
- Let $Z^{(i)} = \{Z^{(i)}(-42), Z^{(i)}(-41), \ldots, Z^{(i)}(27)\}$, $i \in \overline{1, n}$, where $Z^{(i)}(0)$ corresponds to the value of AH' in the day of influenza onset. Form the averaged sample Z: $Z(j) := \frac{\sum_{i=1}^{n} Z^{(i)}(j)}{n}$, $j \in \overline{-42, 27}$.

The sample *Z* contains averages of AH' values which we are going to investigate further to find anomalous drops corresponding to extremely low values of AH'. The simplest way of doing it is to seek them visually on the AH' graphs. However, since the visual comparison of graphs is highly subjective and thus error-prone, it seems reasonable to assess the significance of the observed abnormal drops with the help of statistical tests.

2.2. Statistical testing

In order to determine statistical significance of an AH' deviation from the level $AH' = 0$, a bootstrapping using a Monte Carlo sampling procedure was used similar to the one used in [9]. Our aim was to accept or deny the null hypothesis: samples of AH' dynamics preceding the epidemic onset do not differ significantly from the randomly chosen arbitrary samples of AH' dynamics throughout the year.

To test this hypothesis two samples of AH' are generated. One consists of the averaged values of AH' throughout the wintertime seasons sampled randomly in time and space ("arbitrary sample"), another one is comprised of averaged

AH′ prior to the influenza onset ("preonset sample"). The algorithm aimed at generating these samples is described below.

Arbitrary AH′ *sample generation.* For $i \in \overline{1, 100000}$ do:

- Let *X* be an empty set. Repeat *n* times, where *n* is a total onset events number for a given threshold:
    - Choose randomly a year, a state and a day within the winter season (from October 1 to April 30).
    - Take a sample of *AH*′ values corresponding to the continuous measurements during 4 weeks prior to the selected day for a chosen year and a state (28 *AH*′ values in total).
- Find sample mean and add it to *X*.

The resulting set *X* is the arbitrary sample.

Preonset AH′ *sample generation.* Let *Y* be an empty set, *n* be the total number of registered influenza onsets in different states in different years. For $i \in \overline{1, n}$ do:

- Take a sample of *AH*′ values corresponding to the continuous measurements during 4 weeks prior to the $i - th$ onset (for the corresponding year and a state).
- Find sample mean and add it to *Y*.

The resulting set *Y* is the preonset sample.

Hypothesis testing. To understand whether the samples *X* and *Y* may have the same mean, we use Welch's t-test. This test, apart from the commonly used Student's t-test, does not require equal sample variance. Let null hypothesis be such that samples have identical means. The test measures whether the mean values of the samples differ significantly. In case if *p*-value< 0.05 we reject the null hypothesis, i.e. we state that our preonset sample of *AH*′ significantly differs from the arbitrary sample. That will correspond to the case of anomalous *AH*′ drop. On the other hand, if *p*-value is large, we cannot assume a statistically significant drop in *AH*′ prior to the epidemic onsets.

3. Anomalous AH drops analysis

The algorithm described above was implemented as a set of procedures in Python 3.x programming language using the libraries `numpy`, `matplotlib` and `scipy.stats`. Having created a tool of analysis, we applied it to datasets at hand to examine *AH* drops in various geographical settings.

3.1. Influenza onsets in the USA

To ensure the correctness of our implementation of the algorithm, we decided to start with reproducing the graphs from [9] using the original data kindly provided by Dr. Shaman. Also we extended the original work by analyzing the AH' drops in separate US states.

Data. The *AH* dataset we used contained daily absolute humidity (*AH*) in the US states from 1972 to 2002. Since the ready-made dataset on influenza onset dates was not available, those dates were assessed using the data on weekly pneumonia and influenza mortality (*P&I* mortality). A particular date was considered to be a date of influenza onset in a fixed location (US state) if it belonged to wintertime period and the corresponding *P&I* mortality had been at or above a prescribed threshold level (e.g., 0.01 deaths/100,000 people/day) during the two preceding weeks.

Reproducing the results for the USA as a whole. In figure 1 one can see the dynamics of *AH*′ during six weeks before the onset and four weeks after it averaged by all the *AH*′ datasets for the seasons when the influenza onset was detected. The solid lines of different colors show the results for different *P&I* mortality thresholds (the particular threshold levels are given in legends). The dashed line shows *AH*′ = 0. The number of detected outbreaks obviously depends on the choice of an onset threshold. Depending on the threshold level used, 1185 – 1402 epidemics were detected among
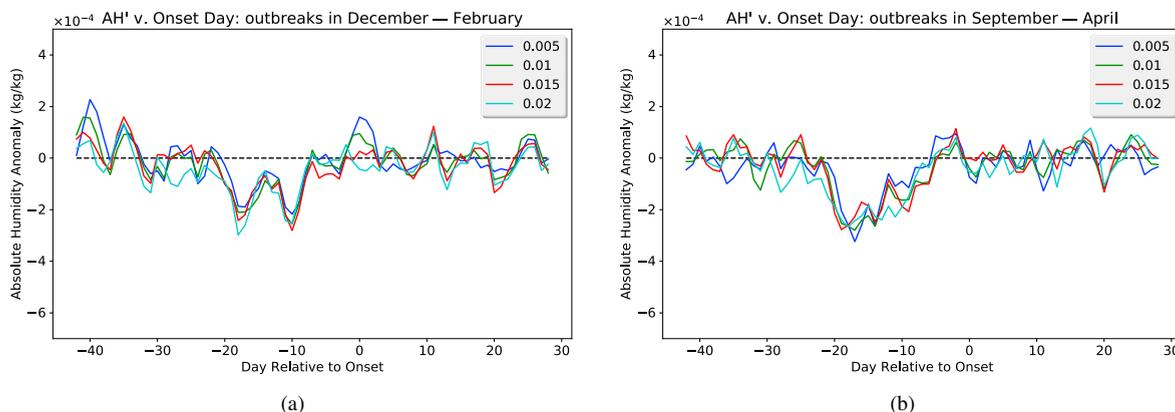
Fig. 1: *AH′* associated with the observed onset of epidemic influenza for two distinct wintertime seasons and four epidemic thresholds.

1470 theoretically possible (30 winters each for the 48 contiguous states plus the District of Columbia). The two graphs in the figure 1 correspond to different ranging of a winter period (i.e. the time period during a season when the *P&I* mortality exceeding the threshold is considered to indicate an onset). Variation of winter period definition changes the total amount of detected outbreaks as well as shifts the onset dates.

The most pronounced *AH′* anomaly is observed in figure 1b when a wintertime period is set to September — April. However, in this case the corresponding overall number of detected epidemics is as large as 1357 – 1467, depending on the thresholds. It seems implausible that influenza epidemics happened almost every winter, thus the correctness of the result may be put under doubt.

We can conclude that the graphs produced by the implemented algorithm are close to the ones demonstrated in [9]. Particularly, for all four winter intervals considered (two graphs were omitted for the sake of saving space), a visually seen *AH′* anomaly is observed at some moment between the day -7 and day -20.

Reproducing the results for state groups. As the averaging of absolute humidity data goes over the whole territory of the USA, it is not fully clear which part of the country gives a greater contribution to the humidity anomaly prior to onset. In [9], the authors addressed this issue by conducting the data analysis separately for four regions of the country: the Southwest (Arizona, Colorado, Nevada, New Mexico, and Utah), the Northeast (Connecticut, the District of Columbia, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and West Virginia); the Gulf region (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, and Virginia), and the rest of the states (21 in overall). Texas and California were excluded from the regional analysis due to their large geographic size and a consequent large variations in AH. The corresponding figures for the mentioned regions obtained by our algorithm are shown in fig. 2a–2d. As it can be seen from the pictures, the distinct *AH′* anomaly before the onset is not detectable for the group of southwestern US states in full accordance with [9]. The results of statistical tests (see table 1) also confirm the original conclusions.

Regarding separate states. It goes without saying that the averaged *AH′* dynamics of a given region is composed of AH dynamics in separate states, and their contribution to the drops of *AH′* is not equal. For example, in Mississippi

Table 1: Welch's t-test results for US state groups, epidemic threshold level 0.02. The results which **do not** correspond to statistically significant *AH′* drops are marked with red.

State group	Southwest	Northeast	Gulf region	The rest
Total onset number, n	132	361	272	571
p-value	0.525	< 0.01	< 0.01	0.022
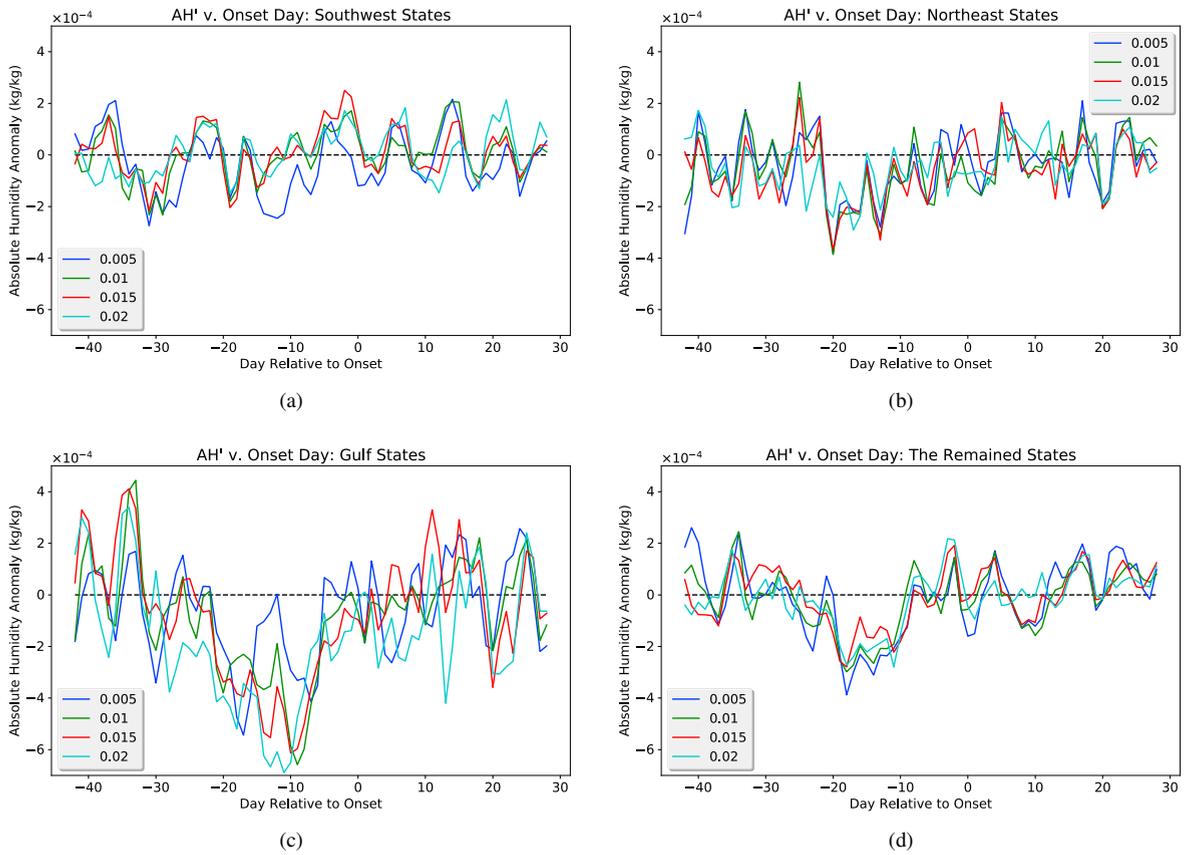
Fig. 2: Averaged *AH′* dynamics for separate US regions
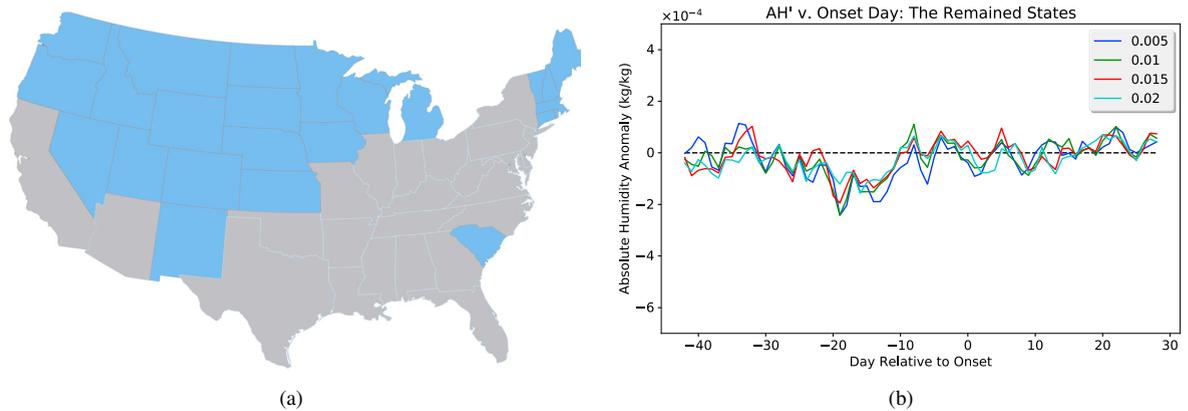
Fig. 3: (a) Grouping of the US states for *k* = 25 (group A is gray, group B is blue); (b) dynamics of *AH′* averaged for the group B states.

Table 2: Welch's t-test results for different sizes k of the group A. The results which **do not** correspond to statistically significant AH' drops are marked with red.

k	0	1	2	3	4	5	6	7	8
p-value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	0.020	0.026	0.028	0.045
k	9	10	11	12	13	14	15	16	17
p-value	0.071	0.142	0.247	0.26216	0.26352	0.26322	0.34268	0.41754	0.52582

AH' drops down to -0.00212 in period from 19 to 10 days prior to onset, which is 7 times greater by absolute value than the minimum AH' drop over all contiguous states. To assess the impact of absolute humidity anomaly in each of the states towards the overall AH' drop, we performed the following data analysis. All the states were sorted by minimal absolute value of AH' detected in the mentioned period. Let's name "group A" the group of k states with most prominent AH' anomaly, whereas the remaining states constitute the group B. If we compare visually the graphs of averaged AH' dynamics in a group B for $k = 0$ (averaging by all the states) and $k = 21$ (the data of 21 states is excluded from the averaging procedure), they contain a distinct and almost similar AH' drop. However, if k reaches 25, the AH' anomaly becomes hardly distinguishable (see figure 3).

To support the visual observations we performed the statistical tests for different values of k. The following procedure was employed:

- Let B be a set of states sorted in ascending order by minimal AH' value within 28-days period prior to onset.
- For each state S in B:
    - Generate arbitrary and preonset AH' samples for the group of states B
    - Perform Welch's t-test to accept or reject the hypothesis of the equivalence of average sample means.
    - Remove S from the group B.

The results demonstrated in table 2 show the growth of $p-$value with the growth of k. One can see that starting from $k = 9$ the null hypothesis of identical sample means is accepted, i.e. the corresponding AH' drop is claimed to be insignificant.

We also performed the same statistical analysis for the data on separate US states. Surprisingly, only three distinct states, – namely, Alabama, Georgia and Louisiana, – demonstrated significant AH' drops.

3.2. Influenza onsets in Russia

Data. The AH dataset contains daily absolute humidity (AH) in the three biggest Russian cities, – Moscow, Saint-Petersburg and Novosibirsk, – from 1985 to 2015. The epidemiological data consists of two datasets. The first one included weekly incidence of acute respiratory infections from June, 1985 to May, 2015 (29 epidemic seasons). The second one, for the same time period, is a binary array marking the outbreak weeks: '1' means officially declared influenza epidemic in the corresponding week, and '0' means the absence of epidemics. This array, along with ARI incidence data, was provided by Russian Research Institute of Influenza [2]. The ARI incidence data was converted to daily one by means of cubic interpolation and corrected for under-reporting during holidays [1], [3].

Onset analysis. To form influenza onset dates array we have used two methods:

- **Using external assessment.** Having taken the binary array as a source of information on epidemic days in a fixed season, we considered Thursday of the first week marked with '1' to be an onset day.
- **Using epidemic thresholds.** Similarly to the rule applied to US *P&I* mortality in the previous section, we consider the particular date to be a date of influenza onset in the particular location if it belonged to wintertime period and the corresponding ARI incidence had been at or above a prescribed threshold level during the two preceding weeks. The threshold levels for ARI incidence were taken in such a way that the total number of detected onsets was close to the number obtained by the external assessment (i.e. the official number of flu epidemics provided by Russian healthcare organs).
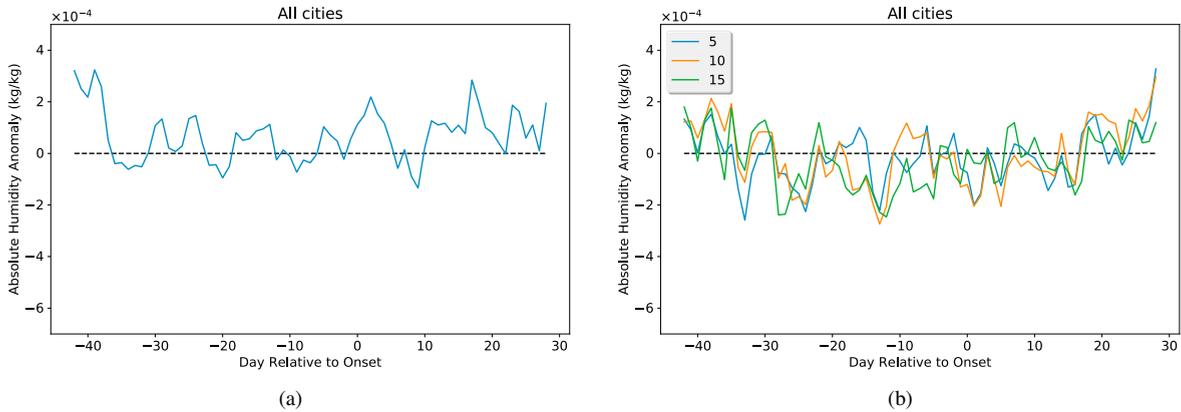
Fig. 4: Dynamics of averaged AH' for three Russian cities with the onsets: (a) taken from external sources; (b) detected using incidence thresholds.

Table 3: Welch's t-test results for Russian cities, the onsets determined by incidence threshold

Threshold	5	10	15	20	25	28	30	35
p-value	0.28639	0.15597	0.05876	0.1501	0.09052	0.03724	0.02619	0.02523
Total onset number, *n*	70	69	62	59	53	51	50	47
Threshold	40	43	44	45	50			
p-value	0.03144	0.0405	0.07254	0.06196	0.14596			
Total onset number, *n*	45	44	43	43	39			

The binary array contained 72 epidemics in total (out of $29 \cdot 3 = 87$ theoretically possible epidemics). The assessed epidemic threshold level appeared to be 5 new flu cases/100,000 people/day for the three cities combined (resulting in 70 detected epidemics). The AH' graphs for the both onset array generation algorithms are presented in figure 4. For both cases there is no visual evidence of AH' drops, and the statistic test proves this fact: for the left graph p-value ≈ 0.78, and for the thresholds on the right graph $p = 0.286$, 0.156 and 0.15 correspondingly (see also table 3).

An interesting fact was revealed when we performed tests with higher threshold levels. Starting from the threshold level 28 we observe a distinct AH' drop before the onsets in the graphs (see figure 5) and the corresponding p-values are below 0.05. A significant AH' drop is demonstrated for the thresholds from 28 to 43. With further increasing of the threshold level the drops again cease to exist.

Table 4: Welch's t-test results for Ile-de-France, depending on morbidity threshold

Threshold	1	2	3	4	5	6	7	8
p-value	0.15874	0.11891	0.08969	0.0382	0.049	0.05086	0.05241	0.09705
Total onset number, *n*	25	25	24	22	21	21	21	21
Threshold	9	10	12	15	17	20	25	30
p-value	0.04526	0.04712	0.02676	0.0209	0.02123	0.04627	0.04445	0.03916
Total onset number, *n*	20	18	17	17	17	15	15	12

3.3. Influenza onsets in Ile-de-France

For the comparison purposes, we decided to test our algorithm on the absolute humidity and incidence data for Ile-de-France (Paris and its suburbs), 1985 – 2015. The French epidemiological data was taken from Sentinelles surveillance network [7]. Apart from the Russian data, it includes only the cases of influenza-like illness (severe ARI forms similar by symptoms to the flu) rather than all the cases of ARI. Epidemic thresholds for the onset detection
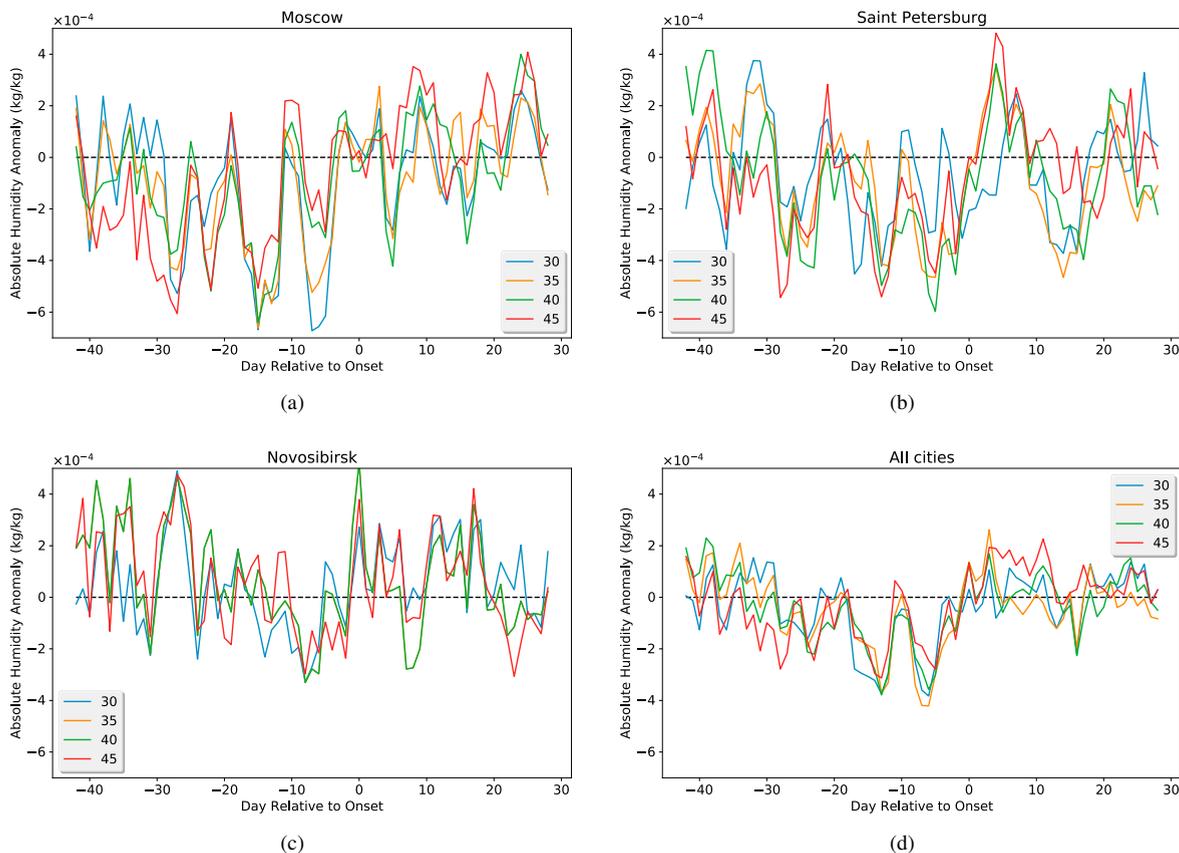
Fig. 5: Dynamics of *AH'* averaged for the 6 wk prior and 4 wk following the onset, high epidemic thresholds
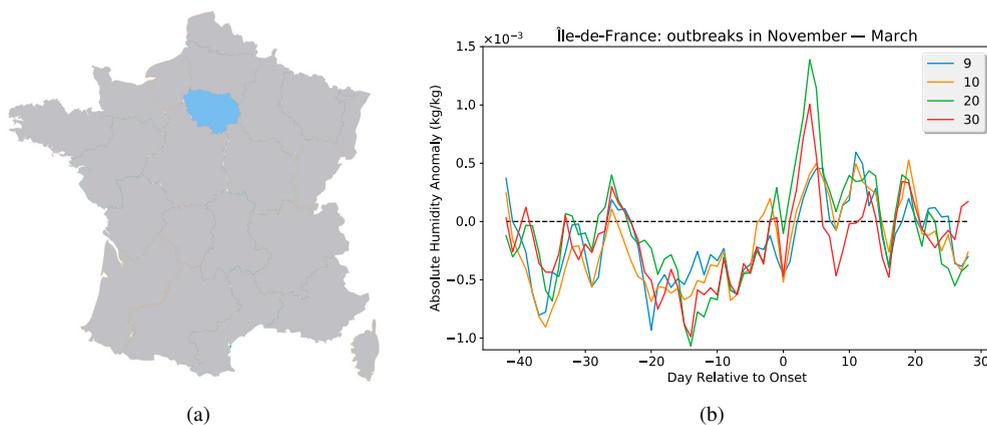
Fig. 6: (a) Ile-de-France region on the map of France; (b) dynamics of *AH'* averaged for the 6 wk prior and 4 wk following the onset, Ile-de-France

algorithms were set somewhat arbitrary to correspond to "reasonable" number of detected epidemics. The wintertime period was assumed to start in November and end in March. The resulting detected onset number n varied from 12 to 20 (of 29 possible in 1985 – 2015 period). The graph for the resulting averaged AH' is demonstrated in figure 6

The first thing one can observe from the graph is that the AH' drop is perfectly visible and, moreover, its absolute value is even greater than of the drops for the US states (see, for instance, figure 2c). Thus, it is not surprising that the statistical tests demonstrate significance of the drops for incidence threshold levels above 9 new incidence cases/100000 people/day (see table 4).

4. Conclusions

In this work we have implemented the computational algorithm which makes it possible to analyze the dynamics of absolute humidity before and after the influenza onsets. Our aim was to find out if the abnormal AH' drops described by Shaman et al. could serve as harbingers of incoming influenza epidemics in the regions outside the USA, namely, France and Russia. Also, the question was, to what extent the results obtained for the US state groups are dependent on the group choice. Our conclusions can be summarized in a following way:

- Regarding the USA, the most distinct AH' drops before the onset happen in the southeastern states, as well as in the states in the south-west of the country that are close to the ocean (Arizona and California). At the same time, the statistical significance of the mentioned drops is reached, according to Welch's t-test, only in three states – the latter result is rather controversial and requires further investigation.
- The Ile-de-France region showed even more distinct drops in AH' than the US states.
- The influenza onsets in Russian cities under consideration are generally not preceded by any abnormal dynamics of AH. A number of cases when we were able to find statistical significance of AH' drops (corresponding to the onset thresholds from 28 to 43) correspond to presumably unrealistic input parameter values (the resulting number of detected epidemics did not exceed 51, which is significantly less than the number of 72 determined by the epidemiologists).

The authors have to admit that the accuracy of the results may be somewhat undermined by the heterogeneity of the data employed (caused by differences in methods of onset detection and big variation in region sizes), uncertainty of the input (namely, the selection of epidemic threshold levels) and the possible limitations of the statistical procedure employed to detect the significant AH' drops. Nevertheless, the current results allow us, somewhat speculatively, to draw a parallel between Russia, France and the US states. The preliminary hypothesis could be made that for the case of locations situated within the latitude span of the US territories (i.e., we exclude the tropics and southern hemisphere), the higher average temperatures (and thus, bigger variation in absolute humidity) and smaller distance to the seashore may lead to more pronounced drops in AH' before the influenza onset. On the contrary, if the region under consideration is located to the north and far from the shore, one won't be successful in predicting influenza onsets by AH' drops. The latter is the case both for the major part of the American North and the three biggest Russian cities.

As a future perspective, we think of applying the algorithm to the extended Russian ARI dataset, particularly, including the Russian cities which are situated closer to the southern seas and have milder climate (e.g., Sochi, Krasnodar, Stavropol). Perhaps, the analysis of such a dataset will bring us to new interesting insights and help specify the conditions under which the AH dynamics may be incorporated into flu prediction models.

Acknowledgements

References

[1] OV Baroyan, UV Basilevsky, VV Ermakov, KD Frank, LA Rvachev, and VA Shashkov. Computer modelling of influenza epidemics for large-scale systems of cities and territories. In *Proc. WHO Symposium on Quantitative Epidemiology, Moscow*, 1970.

[2] Flu Institute. Research Institute of Influenza website. [online]. `http://influenza.spb.ru/en/`.

[3] V. N. Leonenko, S. V. Ivanov, and Yu. K. Novoselova. A computational approach to investigate patterns of acute respiratory illness dynamics in the regions with distinct seasonal climate transitions. *Procedia Computer Science*, 80:2402–2412, 2016.

[4] Vasiliy N Leonenko and Sergey V Ivanov. Influenza peaks prediction in Russian cities: comparing the accuracy of two SEIR models. *Mathematical Biosciences and Engineering*, 15(1):209–232, 2018. DOI: 10.3934/mbe.2018009.

[5] Vasiliy N Leonenko, Yulia K Novoselova, and Karen M Ong. Influenza outbreaks forecasting in Russian cities: Is Baroyan-Rvachev approach still applicable? *Procedia Computer Science*, 101:282–291, 2016.

[6] AA Romanyukha, TE Sannikova, and ID Drynov. The origin of acute respiratory epidemics. *Herald of the Russian Academy of Sciences*, 81(1):31–34, 2011.

[7] Sentinelles. Sentinelles surveillance system website. [online]. `https://websenti.u707.jussieu.fr/sentiweb/`.

[8] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248, 2009.

[9] Jeffrey Shaman, Virginia E Pitzer, Cécile Viboud, Bryan T Grenfell, and Marc Lipsitch. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol*, 8(2):e1000316, 2010.

[10] James Tamerius, Martha I Nelson, Steven Z Zhou, Cécile Viboud, Mark A Miller, and Wladimir J Alonso. Global influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental health perspectives*, 119(4):439, 2011.

[11] Sander P van Noort, Ricardo Águas, Sébastien Ballesteros, and M. Gabriela M Gomes. The role of weather on the relation between influenza and influenza-like illness. *Journal of theoretical biology*, 298:131–137, 2012.

[12] WHO. Influenza (seasonal). Fact sheet No. 211, March 2014. [online]. `http://www.who.int/mediacentre/factsheets/fs211/en/`.