International Journal of Chemical Modeling
Volume 9, Number 1
ISSN: 1941-3955

LOOKING FOR PATTERNS IN EXTREME COLD AND HOT PERIODS ACROSS CANADA AND EURASIA IN DAILY TMAX TEMPERATURES FOR THE LAST 145 YEARS USING 2 STANDARD DEVIATIONS ON EITHER SIDE OF THE MEAN AS REFERENCE POINTS

Darko Butina*

ABSTRACT

This paper has applied one of the gold standards in statistics, the normal distribution bell curve, to identify extreme cold and hot daily temperatures across the Eurasia and Canada covering datasets as old as 1763 up to present day. The normal distribution curve was used to design a classification protocol in which datapoints were partitioned between three classes, an extreme cold class, labelled as e.cold which has the z-scores below -2.0; an extreme hot class, labelled as e.hot, which has z-scores above 2.0; and the major class, labelled as normal with z-scores within +/- 2 standard deviations from the mean. The statistical analysis has been done on the datasets from 19 weather stations at different geographical locations representing Canada and Eurasia countries. The datasets were downloaded, free of charge, from the NCDC/NOAA global daily station data that covers 23,717 different geographical locations across the globe using KNMI Climate Explorer to access the data. Systematic statistical analysis was performed on close to 1 million daily tmax thermometer readings, ranging from temperatures as low as -50.0C and as high as 45.0C, with a total range of 95.0C.

The main conclusion of the paper is that air temperatures across Eurasia and Canada are dominated by extreme cold temperatures, while extreme hot temperatures cannot be differentiated from the normal variations which are within +/- 2 standard deviations from the mean. This conclusion is based on the fact that every single weather station tells the same story - each weather station's distribution curve is heavily skewed to the left of the mean, i.e., to the cold tail of the curve.

---

* *Dr Darko Butina is a retired scientist with 40 years of working in experimental and computational fields of drug discovery. For any enquiry please contact the author at* darko.butina@chemomine.co.uk.

INTRODUCTION

In the previous paper by this author [1], a new statistical protocol for partitioning daily tmax temperatures into three classes, normal, extreme hot and extreme cold, has been developed and used to analyse temperatures across the USA. The protocol is based on the concept of the normal distribution curve by which all datapoints that lie between 2 standard deviations on either side of the mean are labelled as ***normal***, while those that lie outside 2 standard deviations on either side of the mean are labelled as ***extreme*** [2,3 and 4]:

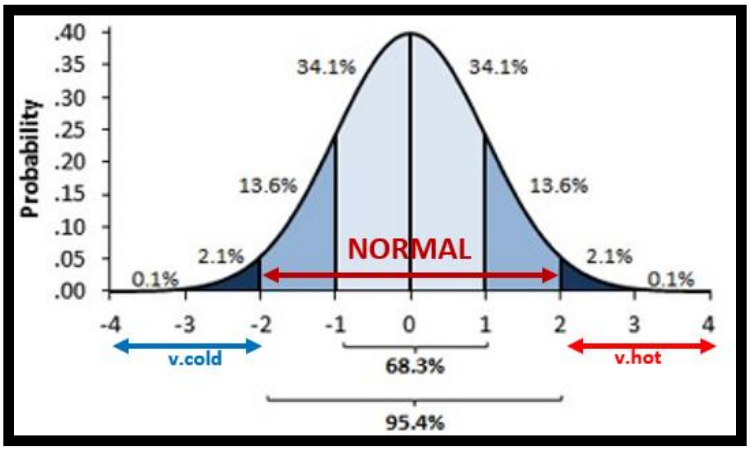

Figure 1. Normal distribution bell-shaped curve where 95% of data reside within +/- 2 standard deviations from the mean and are labelled as normal, while 5% of the datapoints which are outside that range will be labelled as extreme.

The key points of Figure 1 are that approximately 95% of the datapoints are expected to be within 2 standard deviations around the mean, while the remaining 5% of the data would be outside 2 standard deviations from the mean, 2.5% outside -2 standard deviations (to the left) of the mean and 2.5% outside (+)2 standard deviations (to the right) from the mean. Since the *distance from the mean in standard deviations is known as* the ***z-score,*** and since the datasets that will be analysed are generated by a calibrated thermometer, the classification protocol is formulated in the following:

Normal (between points A and B):	-2.0 >= z-score <= 2.0	(1)
e.cold:	z-score < -2	(2)
e.hot:	z-score > 2	(3)

Figure 2. Three-class protocol that will be used throughout this paper.

The label *e.cold* is short for *extreme cold*, while *e.hot* is short for *extreme hot*. Combining Figure 2 with the labels normal, e.cold and e.hot a simplified classification scheme has been designed:

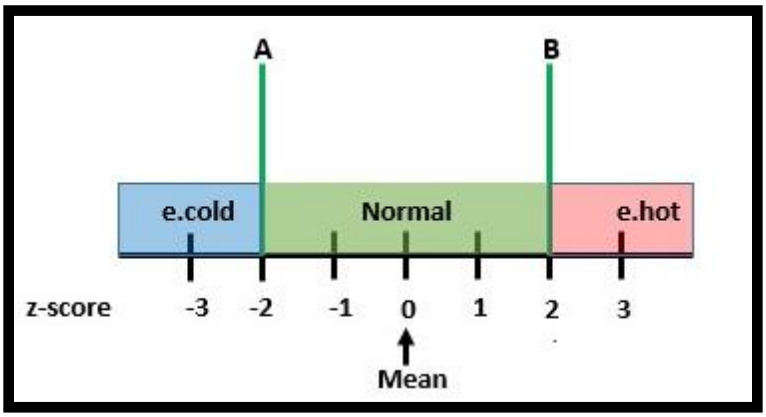

Figure 3. Classification rules to assign each datapoint into one of the three classes, **e.cold**, **normal** or **e.hot.**

The points A and B (Figure 3) represent ***normality boundaries*** and all the datapoints **within**, or inside, those two boundaries will be assigned the label ***normal***. Datapoints that are outside the lower normality boundary A (A has z-score = -2.0) will be labelled as e.cold, while those above the upper boundary B (B has z-score = 2.0) will be labelled as e.hot.

So, the datapoint with z-score = -3.4 will be labelled as e.cold, datapoint with z-score = 1.2 will be labelled as normal while the datapoint with z-score = 2.1 will be labelled as e.hot.

To transform any given datapoint to a z-score the following formulae is used:

$$z\text{-}score = (X\text{-}Mean)/SD \quad \text{(F1)}$$

where **X** is the original datapoint, while **Mean** and **SD** are the mean and the standard deviation for the dataset used. In excel, the mean and the standard deviation can be calculated by using functions ***average*** and ***stdevp*** respectively.

One of the main features of the use of z-scores that cannot be emphasised enough, is the ability to 'back-transform' z-score to the datapoint's original value:

$$X = Mean + (z\text{-}score * SD) \quad \text{(F2)}$$

For example, if a dataset has the mean = 15.0, SD = 10.0 and datapoint X = 20.0C, the z-score would be:

z-score = (20.0–15.0)/10.0 = 0.5 using formula F1.

However, if we want to find the original datapoint for, say, z-score = 2.1 using the same mean and SD as above, **X = 15.0+(2.1*10.0) = 36.0C** using formula F2.

Now that all the definitions and the classification protocol have been established let us move to the issue of the datasets that will be analysed.

DATASETS

The paper will be analysing datasets from 19 weather stations which are based across Canada and Eurasia but excluding the USA which was part of the previous paper [1]. Each weather station's dataset contains maximum daytime air temperatures, tmax, for at least 100 years and also one of the oldest weather station in Milan, Italy, with 246 years of recorded daytime temperatures!

All the data is available free of charge from the NCDC/NOAA global daily station data that covers 23717 different geographical locations. The data can be accessed for download via KNMI Climate Explorer (http://climexp.knmi.nl/start.cgi)

The datasets were selected based on two simple criteria:

- The weather stations should be representative of Canada and Eurasia
- Each weather station should have at least 100 years of data

A few words are needed here about the history of collecting daily temperature data and the importance of the calibrated thermometer. The first systematic recording of the maximum and minimum daily temperatures across the globe started in the mid-1800s, where the highest daily temperature was labelled as tmax while the lowest as tmin. The observations protocol designed by the meteorologists of that time makes scientific sense for the following reasons:

- the highest temperature during daytime, tmax, corresponds to the kinetic energy of the air molecules that surround the thermometer. That in turn tells us about the amount of the heat energy generated by the sun that has reached the fixed-to-the-ground thermometer. *In the thermodynamic terms, the tmax reflects the process of warming*
- on the other hand, tmin observed during night time, when the sun does not heat that part of our planet, *reflects the process of cooling*, thermodynamically a very different process from the warming process that is captured by tmax

In order to make sense of any temperature patterns generated during the day (tmax) and those generated during the night (tmin), we need to separately treat tmax data from tmin data.

In this paper, only maximum daytime temperatures will be analysed.

Format and the Size of Datasets

All the downloads from KNMI Climate Explorer consist of 4 columns that are labelled here as yy (year), mm (months), dd (day) and tmax (daytime maximum temperatures in degrees C).

The first column labelled '*order'* was created by the author to help with sorting and allowing re-setting the dataset to its original order.

It is very important that any sorting on any individual column must be done on all columns since the internal relationship between each column must be maintained throughout the analysis.

Table 1. Typical table that comes from download of tmax data using KNMI Climate Explorer

order	yy	mm	dd	tmax
1	1893	1	1	-2.8
2	1893	1	2	-15
3	1893	1	3	-8.9
4	1893	1	4	1.7
5	1893	1	5	-13.3
	.....	.....	.....	.....
43762	2016	9	29	20
43763	2016	9	30	21.1
43764	2016	10	1	18.3
43765	2016	10	2	24.4

Daily Tmax Temperature Patterns for 19 Weathers Stations across Canada and Eurasia

WS in WS-Code and WS Name stands for Weather Station, # in Years and datapoints stands for 'number of', SD stands for Standard Deviation, Max and Min indicate maximum and minimum daily temperature observed at each weather station, while Total Range indicates a total temperature range observed at a given weather station obtained by taking difference between maximum and minimum daily temperature, tmax.

Table 2. Summary table for 19 weather stations across USA and Eurasia with the historical data as early as 1763

WS-Code	WS Name	Country	Years	# Years	# datapoints	mean	SD	Max Tmax	Min Tmax	Total Range
WS-1	Milan	Ita	1763-2008	246	89259	17.0	9.5	41.1	-9.3	50.4
WS-2	Prague	Czech	1775-2005	231	84103	13.0	9.4	37.8	-21.5	59.3
WS-3	Bologna	Ita	1814-2003	190	69242	17.5	9.5	39.8	-8.8	48.6
WS-4	UCCLE	Bel	1833-2016	184	65179	14.1	7.7	38.8	-13.6	52.4
WS-5	Toronto	Can	1840-2003	164	59648	12.3	11.1	40.6	-25.0	65.6
WS-6	Zagreb	Cro	1881-2016	136	49268	15.7	9.6	40.3	-15.5	55.8
WS-7	Berlin	Ger	1876-2016	141	49297	13.1	9.0	37.7	-16.4	54.1
WS-8	Karlsruhe	Ger	1876-2008	133	48145	14.7	8.9	40.2	-14.0	54.2
WS-9	Kiev	Ukr	1881-2016	136	47617	12.0	11.5	39.9	-25.6	65.5
WS-10	Tashkent	Uzb	1881-2016	136	45976	20.9	11.7	44.6	-18.5	63.1
WS-11	Cagary	Can	1881-2012	136	46849	10.3	12.3	36.1	-38.9	75.0
WS-12	Moscow	Russ	1893-2016	124	44128	14.7	10.6	42.8	-22.8	65.6
WS-13	Montreal	Can	1871-1993	123	44094	10.7	12.5	36.1	-28.9	65.0
WS-14	Osijek	Cro	1899-2014	116	41712	16.6	10.2	40.3	-15.6	55.9
WS-15	Marseille	Fra	1897-2003	107	38650	19.6	6.8	40.6	-6.0	46.6
WS-16	Taganrog	Russ	1916-2016	101	34816	13.9	11.8	40.5	-25.0	65.5
WS-17	Enisek	Russ	1887-2016	130	33128	4.6	15.8	35.9	-49.1	85.0
WS-18	Mokpo	Kor	1926-2016	91	32696	18.4	9.0	37.0	-5.9	42.9
WS-19	Krasnyj	Russ	1913-2016	104	29094	5.8	16.5	39.3	-38.8	78.1
			Mean	**144**	**50153**	**13.9**	**10.7**	**39.4**	**-21.0**	**60.5**

The key reference points *for each weather station*, needed to perform this analysis are:

- number of datapoints
- the mean and the standard deviation
- minimum and maximum observed daily tmax temperature

To demonstrate the whole process of generating the relevant numbers and subsequent analysis, one of the oldest weather station in the North Hemisphere, at Milan, Italy, was chosen, where the daily temperature readings have started back in **1763.**

The analysis starts with the small table consisting of 5 key reference points:

Table 3. Starting table for WS-1 (Milan, Italy)

WS-Code	# datapoints	mean	SD	Max Tmax	Min Tmax
WS-1	89259	17.0	9.5	41.1	-9.3

The first step in this classification protocol is to transform each of the 89,259 temperature readings into their z-scores using ***z-score = (X-Mean)/SD*** formula, where **X** is the original datapoint, while **Mean** and **SD** are the mean and the standard deviation for the dataset used. In excel, the mean and the standard deviation can be calculated by using functions ***average*** and ***stdevp*** respectively.

Table 4. Additional reference numbers that are needed to produce statistical summary for WS-1

WS-Code	# datapoints	mean	SD	Min Tmax	z-score-Min	A (z-score=-2.0)	B (z-score=2.0)	Max Tmax	z-score Max
WS-1	89259	17.0	9.5	-9.3	-2.8	-1.9	35.9	41.1	2.5

The key reference points that define WS-1 are:

- Minimum observed temperature, Min Tmax, is -9.3C with z-score = -2.8
- The lower normality boundary A (see Figure 3) has value of -1.9C
- The upper normality boundary B (see Figure 3) has value of 35.9C
- Maximum observed temperature, Max Tmax, is 41.1C with z-score = 2.5

So, every datapoint with a z-score less than -2.0 will be assigned label ***e.cold***, those with a z-score larger than 2.0 will be assigned label ***e.hot***, while the rest of datapoints that are between the normality boundaries A and B will be assigned label ***normal***.

Since there is a simple relationship between z-scores and the original temperature readings, we can also express those ranges in temperature terms – all datapoints between -1.9C and 35.9C will be classified as normal, those below -1.9C as e.cold and those above 35.9C as e.hot.

Two indices which are instrumental in making the final assessments of the extreme temperature patterns have been identified in the previous paper [1] which are the ***frequency and*** the ***range*** of each class.

Table 5. Frequency of 3 classes for WS-1

	#	%
e.cold	394	0.44
e.hot	204	0.23
normal	88661	99.33

In terms of frequency, 99.3% of datapoints, class normal, are within +/- 2 standard deviations from the mean, 0.4% are in the extreme cold region and 0.2% in the extreme hot region. So, in terms of frequency between e.cold and e.hot class, it is the e.cold class that dominates the e.hot class by 2:1. However, notice that 99.3% of datapoints are within +/- 2.0 standard deviations and therefore classified as normal.

In terms of the range of z-scores that are outside the normality boundaries A and B (see Figure 3), the e.cold class extends to z-score = -2.8, while the e.hot to z-score = 2.5, in other words, the extreme cold class extends 0.3 standard deviations deeper into the extreme tail of distribution curve.

Frequency and Ranges, Based on z-Scores, for the Three Classes across Canada and the Eurasia

The datasets that fit the ideal distribution bell-shaped curve are expected to have approximately 95% of datapoints within +/- 2 SD on either side of the mean and 5% in the extreme part of the curve's tail equally distributed as 2.5% on the opposite sides of the curve.

Table 6. Frequency of three classes across Eurasia

WS-Code	# datapoints	% Normal	% e.cold	% e.hot	Ratio e.cold/e.hot
WS-1	89259	99.33	0.44	0.23	1.9
WS-2	84103	97.97	1.49	0.54	2.8
WS-3	69242	99.30	0.47	0.23	2.0
WS-4	65179	97.16	1.55	1.30	1.2
WS-5	59648	98.39	1.42	0.19	7.5
WS-6	49268	98.28	1.48	0.25	5.9
WS-7	49297	97.59	1.42	0.99	1.4
WS-8	48145	97.69	1.31	0.99	1.3
WS-9	47617	98.39	1.51	0.11	13.7
WS-10	45976	98.17	1.82	0.01	182.0
WS-11	46849	95.22	4.75	0.03	158.3
WS-12	44128	98.14	0.65	1.20	0.5
WS-13	44094	97.83	2.17	0.01	217.0
WS-14	41712	98.31	1.52	0.16	9.5
WS-15	38650	97.90	1.39	0.71	2.0
WS-16	34816	98.62	1.30	0.07	18.6
WS-17	33128	97.40	2.60	0.01	260.0
WS-18	32696	98.62	1.35	0.03	45.0
WS-19	29094	98.68	1.32	0.01	132.0
	Average	98.0	1.6	0.4	62.2

As it can be seen from Table 6, the temperature-based datasets have at least **95.2%** of datapoints in class normal at WS-11, and as much as **99.93%** of datapoints at WS-1. On average, the frequency for class normal for Canada and Eurasia is 98% (last row, Table 6) which in turn means that only 2% of datapoints can be classified as either e.cold or e.hot.

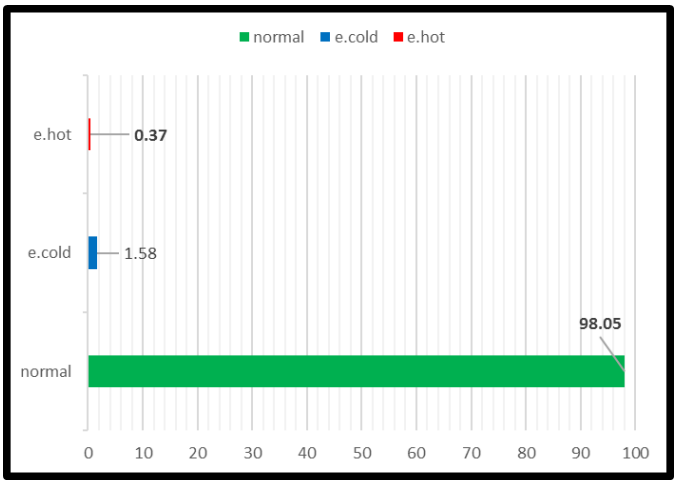

Figure 4. An average distribution, in % of total, for 3 classes across 19 weather stations.

The average ratio of extreme cold and extreme hot datapoints is 81% vs 19%, i.e., for every 100 datapoints in extreme region of the distribution curve, 81 would classified as e.cold while 19 are classified as e.hot:

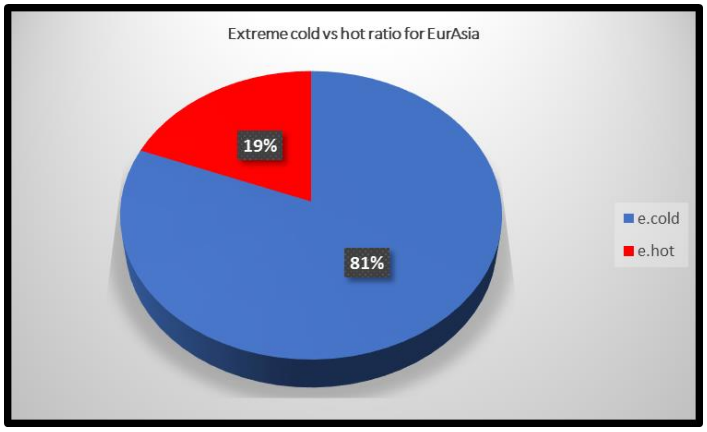

Figure 5. An average ratio of e.cold vs e.hot class for 19 weather stations.

In terms of z-score ranges, the average minimum value for z-score is 3.3 standard deviations below the mean, while the average maximum value is 2.4 standard deviation above the mean:

Figure 6. The average z-score range for e.cold class (blue) and e.hot class (red).

The figure below is a summary of the frequency and the ranges averages for the three classes for Canada and Eurasia weather stations:

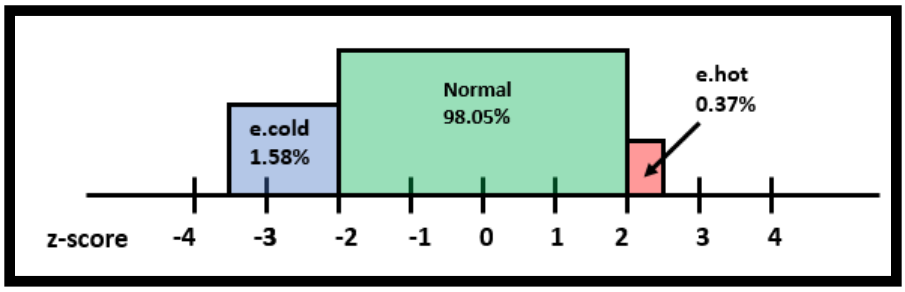

Figure 6. An average frequencies and the ranges for 3 classes across 19 weather stations.

Temperature Ranges for the Three Classes across Canada and the Eurasia

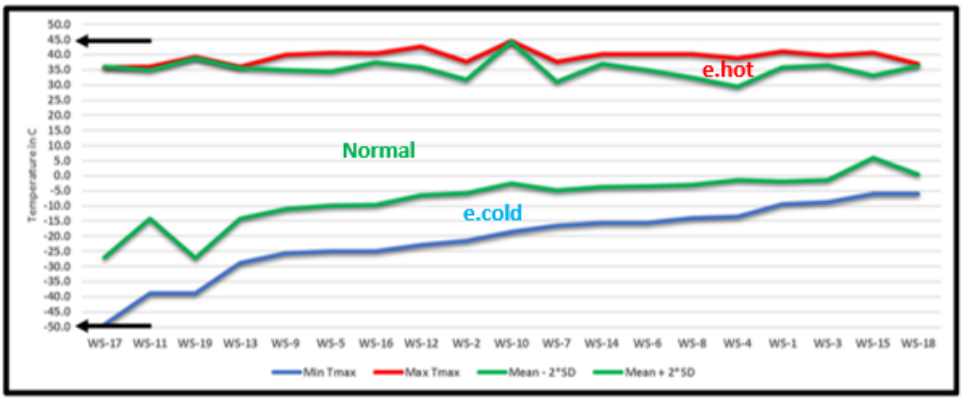

Figure 7. The maximum (red), minimum (blue) and upper/lower normality boundaries (green) for 19 weather stations sorted on the minimum observed temperatures.

The main points in Figure 7 are that the total range of recorded temperature is about 95.0C, with lowest recorded temperature at -50.0C and the highest recorded at 45.0C. While it is very difficult to differentiate the hottest temperatures from the upper normality boundary, at the top of the graph, there is a very clear separation between the lower normality boundary and the coldest recorded temperatures at the bottom of Figure 7.

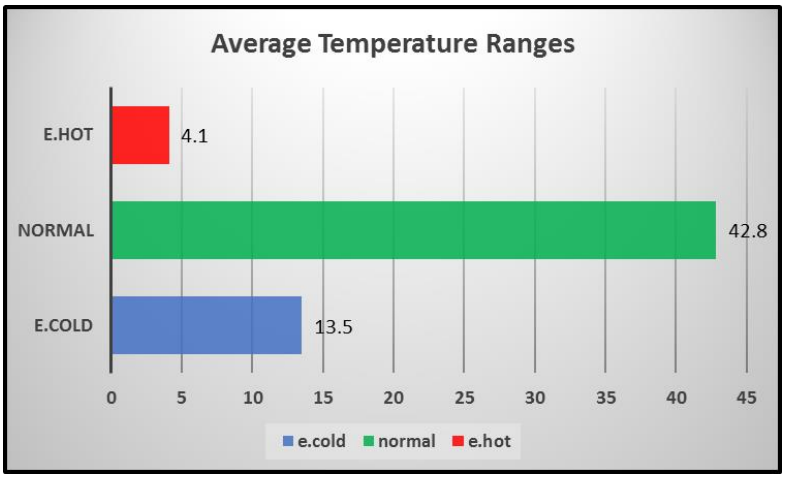

Figure 8. Average temperature ranges for 3 classes.

As can be seen from Figure 8, the average temperature range for class normal is 42.8C, for class e.cold 13.5C while only 4.1C for class e.hot.

While this paper's starting point has been the concept of the symmetrical distribution curve, the science of statistics tells us that any given dataset can be best described by one of the three distribution curve shapes:

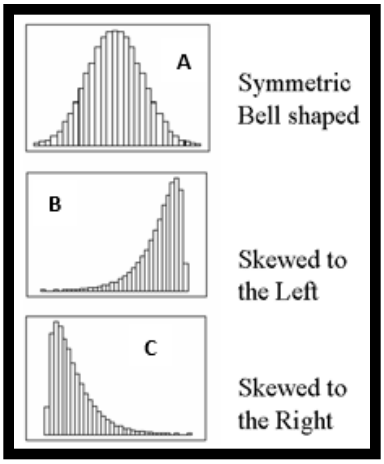

Figure 9. Three basic types of the distribution curves: Symmetric (A), Skewed to the left (B) and Skewed to the right (C).

As we have seen from our analysis, every single weather station can be best described by the shape B (Figure 9), i.e., skewed to the left, or in our case skewed to the extreme cold part of the curve.

SUMMARY AND CONCLUSION

This paper has applied one of the gold standards in statistics, the normal distribution bell-shaped curve, to identify the extreme cold and hot daily temperatures across Canada and Eurasia. The concept of the normal distribution curve was used to design a classification protocol in which datapoints were partitioned between 3 classes:

- e.cold class which was defined by z-score < -2.0
- e.hot class defined by z-score > 2.0 and
- normal class defined by z-score > = -2.0 and < = 2.0

Almost 1 million maximum daily temperatures were analysed distributed across 19 weather stations spread across Europe and Eurasia and with a total range of temperatures of 95.0C.

The paper did not start by trying to prove or disprove any model or hypothesis published in different fields of the climate sciences, but simply perform a systematic statistical analysis of original records that are in the public domain.

Since the paper uses a clearly defined classification protocol and the original data without any modifications, and most importantly, data that is in the public domain and freely available to download, all the results and conclusions can be easily validated by any scientist *who is familiar with the statistical tools used in this paper and understands the physical meaning of the numbers that are generated by a calibrated thermometer.*

In conclusion:

- every weather station's distribution is significantly skewed to the extreme cold tail of the distribution curve (Fig.9, shape B)
- on average, the temperature range for e.cold class is 13.5C while temperature range for the e.hot class is 4.1C (Fig.8)
- in percentage terms, e.cold class covers on average 1.58% of datapoints, the e.hot class only 0.37% while the class normal covers on average 98.05% of the total
- the overall trend for the temperatures in the Eurasia and Canada over the last 145 years, in terms of the tails distribution, is that the e.cold class dominates e.hot class by 4.3 to 1

NOTE ABOUT THE AUTHOR

Dr Darko Butina is a retired scientist with 20 years of experience in experimental organic and medicinal chemistry plus a further 20 years of working in the field of pattern recognition and datamining of experimental data. He was part of the team that designed and synthesised the first effective drug, Sumatriptan, for treatment of migraine – an achievement for which the team at Glaxo received The Queens Award. For more than twenty years Sumatriptan has improved the quality of life for millions of migraine sufferers worldwide. While working in

computational drug discovery, the author developed a novel clustering algorithm, dbclus, that became the de facto standard for quantifying diversity in world of molecular structures.

Since his retirement, he applied various numerical tools developed in fields of pattern recognition, datamining, machine learning to mention but a few, to analysis of the thermometer-generated data and has published 5 papers since 2012 [1,5,6,7and 8]. He also runs his own website at www.l4patterns.com.

ACKNOWLEDGMENT

I acknowledge KNMI Climate Explorer for free access to the global temperatures database which was used in this paper. I also acknowledge my wife, Judy Glasman for proof reading this paper and moral support.

REFERENCES

[1] Butina D., Looking for patterns in extreme cold and hot periods across the USA in daily tmax temperatures for last 125 years using 2 standard deviations on either side of the mean as a reference points, *Int. J. of Chemical Modeling*., 2017, 9, Number 1, 1-15.

[2] Altman D., Why We Need Confidence Intervals, *World J. Surg*., 2005, 29, 554–556.

[3] Copeland B., A practical means for measurement and control of the precision of clinical laboratory determinations, *Precision in Clinical Laboratories*., 1957, 553, 551-558.

[4] Finney D., *Experimental Design and its Statistical Basis,* University of Chicago Press, 1955, p. 169.

[5] Butina D., Should we worry about the Earth calculated warming at 0.7C over last 100 years?, *Int. J. of Chemical Modeling*., 2012, 4, Number 2-3, 233-253.

[6] Butina D., Quantifying the effect that N2, O2 and H2O have on night-to-day warming trends at ground level, *Int. J. of Chemical Modeling*., 2013, 5, Number 4, 457-478.

[7] Butina D., Is Arctic Melting? Theory vs. Observations, *Int. J. of Chemical Modeling*., 2015, 7, Number 1, 91-113.

[8] Butina D., New algorithm to identify coldest and hottest time periods. Case study: Coldest winters recorded at Armagh Observatory over 161 years between 1844 and 2004, *Int. J. of Chemical Modeling*., 2015, 7, Number 3, 212-227.