Identification of buffered data in time series preprocessing: application on surface river temperature

Nelly Moulin, Frederic Gresselin, Bruno Dardaillon, Zahra Thomas

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Identification of buffered data in time series preprocessing: application on surface river temperature

Nelly Moulin[1] | **Frederic Gresselin**[2] | **Bruno Dardaillon**[3] | **Zahra Thomas**[1]

[1]UMR 1069 SAS, Institut Agro, Bretagne, France

[2]UMR 6143 M2C, Université de Caen-Normandie, Normandie, France

[3]UMR 6139 LMNO, Université de Caen-Normandie, Normandie, France

Correspondence
Corresponding author Nelly Moulin,
Email: nelly.moulin5@gmail.com

Present address
65 rue de St Brieuc 35042 Rennes Cedex, France

Abstract

With the growing number of sensors technologies, the production of numerous types of data allows finer observations of our environment. Among them, time series represent a valuable heritage by the time spent on their recording and the information they contain. However, the analysis of time series produced by a monitoring network generally requires preprocessing steps to separate data with meaningful information from sensors' dysfunctions or measurement particular conditions. In this context, outliers are already well studied and several methods are already developed to identify them. In this paper, we propose a complementary method to identify buffered data. Buffered data are characterized by a lower amplitude than the rest of the time series and can be naturally caused (groundwater influence for example) or caused by measurement defects (sensor covered by sediment movements). The necessity to identify buffered signals came with the use of data coming from several databases with different level of qualification. Buffered signals are not necessarily filtered with conventional preprocessing methods and can affect the analysis when not related to the studied phenomena. The identification method proposed in this study relies on a normalized diurnal range index. It was developed on surface river temperature time series recorded in metropolitan France to cover a wide variety of regional climates and measurement environments. The method is able to highlight buffered data inside a time series. Furthermore, it is able to separate (naturally caused or not) occasional or regular buffered signal periods in a time series. The study then uses preprocessed time series to analyze the distribution of regular buffered data according to the season of occurrence and a climate typology.

KEYWORDS
data processing, thermal regime, diurnal temperature range index, time series, classification, buffered data, physical variables, water temperature

1 | INTRODUCTION

In a context of climate change, our temporal understanding of the environment can benefit from numerous time series. Their interpretation relies on the study of different frequency signals disturbed by noise or singularities[1]. However, not all singularities are to be withdrawn and a number of them contain valuable information[2].

In sciences dealing with environmental systems such as hydrology, recent climate changes cause numerous singularities. In order to anticipate necessary adaptations and management strategies, understanding how our environment reacts to climate change becomes more important every day[3,4,5,6]. In this context, identifying singularities produced in changing environments is essential. Streams and rivers are among the ecosystems most sensitive to climate change. Thus, the evolution of their temperature (from here on referred to as "water temperature" T_W) is a key parameter that is commonly monitored[7,8,9,10,11]. Most of the time, it is measured at punctual locations. However, several events can alter the quality of the measurement depending on the environmental configuration[12]. For instance, streams can easily be subject to low water level leading to unusual water temperature signals[13], artificial facilities such as dams can also lead to sensors covered by sediment and thermal changes[14,15,16,17]. Reading a water temperature time series and sorting the different events can improve its use and help characterise the thermal

Abbreviations: Ta, air temperature; Tw, water temperature; TS, time series.

regime of the river at the measurement location[18,19].

This paper focuses on the identification of one type of singularity in time series: signals showing reduced amplitude variations (from here on referred to as "buffered signal") compared to the rest of the time series or to other time series in the same geographical context. This reduced amplitude can be naturally caused, such as groundwater inflows or humanly caused, such as sensors moved in the river. This type of singularity is not as common as outliers and its effects are less visible than extreme data so fewer methods were developed to identify it. The identification method is applied directly on raw temperature time series[20,21,22,1,23]. It does not require any filtering or smoothing steps. Therefore, it prevents additional effects and loss of information caused by these common preprocessing methods[24,25]. The method relies on a normalized diurnal range index which is calculated with the time series itself.

The application of the method to the whole dataset confirmed the efficiency of the identification method to 1. identify buffered data periods in time series and 2. distinguish buffered signals being part of the river thermal regime from buffered signals caused by extrinsic factors. Further analysis partly explains the distribution of naturally caused buffered signals with the season of occurrence and a classification of regional climate types.

2 | MATERIAL AND METHODS

2.1 | Studied area

The studied area is $552.000\,\mathrm{km}^2$ and corresponds to metropolitan France (Fig. 1). The territory is characterized by oceanic climates on the west part along the Atlantic ocean. The east, center and south west parts are marked by old and young mountains and the highest annual precipitations rates. The mean water temperatures vary from 11 to $14\,^{\circ}\mathrm{C}$ in the northern half to more than $17\,^{\circ}\mathrm{C}$ in the southern half with lower mean water temperatures in mountain areas in the Alps (East frontier), in the Massif Central (Center) and in the Pyrenees (South West frontier). In this study, we used a classification in 8 types of climate for metropolitan France determined by[26] from a combination of 14 variables. Time series used in this study cover each type of climate.

- Type 1: Mountain climate characterized by a high rate of precipitation, a mean air temperature less than $9.4\,^{\circ}\mathrm{C}$, more than 25 cold days (i.e. with a minimum temperature less than $-5\,^{\circ}\mathrm{C}$) and less than 4 hot days (i.e. with a maximum temperature above $30\,^{\circ}\mathrm{C}$).
- Type 2: Semi-continental climate, located at the border of Type 1 climate zones. Air temperatures are slightly higher than in Type zones but colder than other regions at the same elevation. The precipitation rate is also slightly lower than in Type 1 zones but with a similar climate variability. The fall precipitation rate over the summer precipitation rate is particularly low for this climate type.
- Type 3: Degraded oceanic climate for Center and North areas. The mean temperature is intermediate ($11\,^{\circ}\mathrm{C}$) and the precipitation rate is low (less than 700 mm annually). The precipitation variability is very low whereas the temperature variability is very high.
- Type 4: Altered oceanic climate, located between Type 3 and Type 5 climates. The mean temperature is higher ($12.5\,^{\circ}\mathrm{C}$). The number of cold days is low (4 to 8 per year) whereas the number of hot days is high (15 to 23 per year). The annual temperature amplitude is the smallest and the precipitation rate occurs mostly in winter (summers are quite dry) with an annual accumulation of $800\,\mathrm{mm}$.
- Type 5: Frank oceanic climate, along the Atlantic ocean and the British channel. The annual thermal amplitude is quite low (less than $13\,^{\circ}\mathrm{C}$) as well as the number of cold days and hot days (less than 4 per year for both). The precipitation rate is quite high (more than $1000\,\mathrm{mm}$) and presents a high inter-annual variability in winter.
- Type 6: Altered Mediterranean climate located mainly in the South Alps. It is characterized by a high annual mean temperature with a reduced number of cold days and a high number of hot days (between 15 and 23 per year). The inter-annual variability of July temperatures is minimum among the climate types. The annual precipitation rate is intermediate (800 to $900\,\mathrm{mm}$/year but mostly distributed in fall and winter, leaving dry summers.

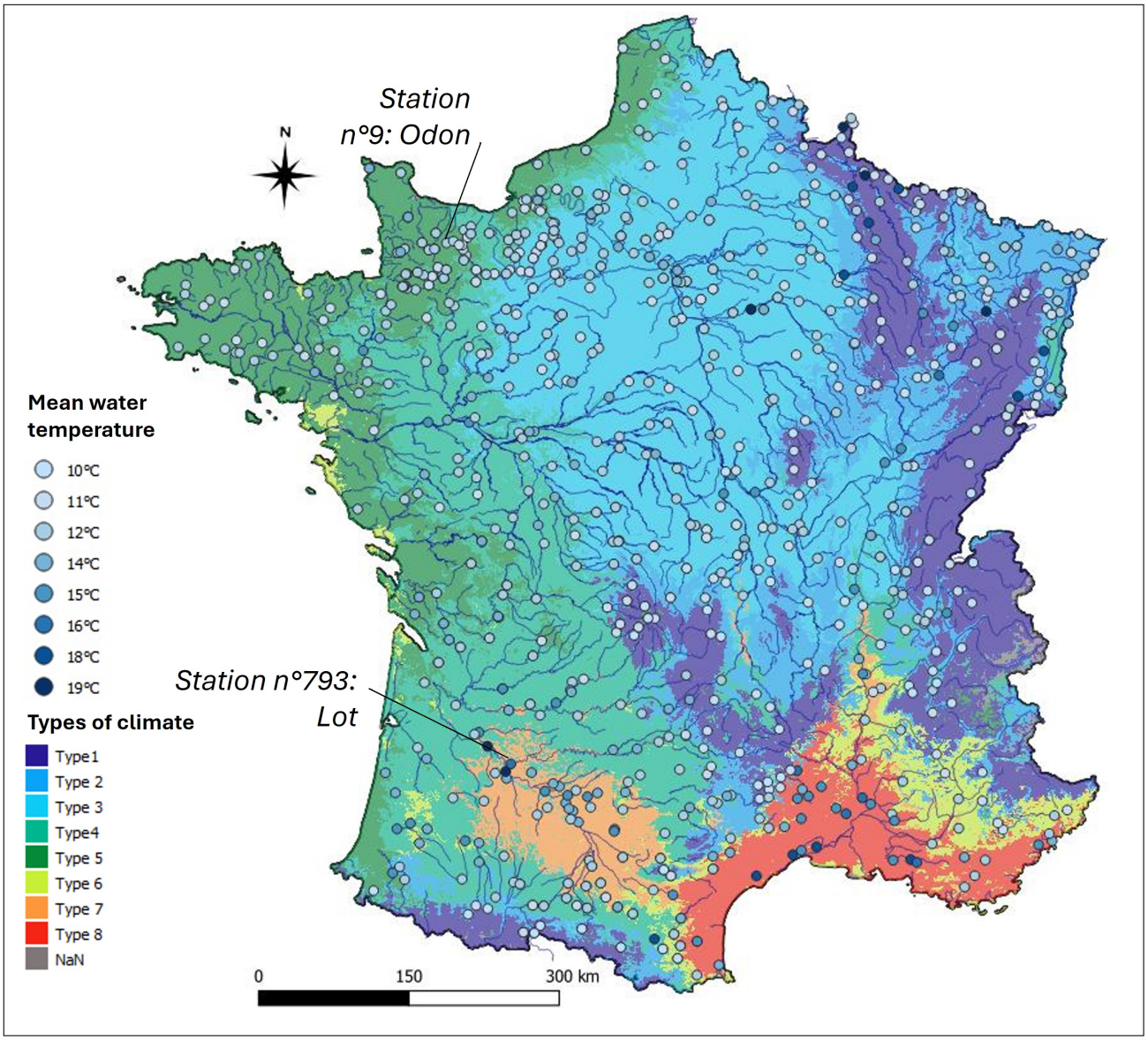


FIGURE 1 Distribution of the 993 measuring stations. Blue coloured dots correspond to the mean water temperature (in °C) of the collected time series. Background map indicates the typology of climates defined by[26]. The "NaN" category corresponds to zones with no specific climate type attributed.

- Type 7: South-West basin climate located mainly in the *Garonne* catchment. The annual temperature is high (above 13 °C) and a high number of hot days (above 23 per year). The annual thermal amplitude is high (15 to 16 °C) with a low inter-annual variability. The precipitation rate is low (less than 800 mm) and more frequent in winter (scarce rainfalls) than summer (strong rainfalls in less than 6 days).
- Type 8: Frank Mediterranean climate. Characteristics are very contrasted with a high annual temperatures and a high annual amplitude (more than 17 °C between July and January) repeatable from one year to another. The fall precipitation rate over the summer precipitation rate is very high (> 6).

2.2 | Water temperature time series

Three water temperatures databases were used in this study in order to cover the largest area, gather the most cases of signal amplitude and cover the different climatic zones. In total, the dataset contains water temperature time series from 993 measurement stations spread on the whole territory. The three databases are maintained by three different institutions (Fig.2).

- DREAL database: this network is held by the Direction Régionale de l'Aménagement et du Logement (DREAL) of Normandy. It is composed of 31 HOBOware sensors with an accuracy of $0.2\,°C$ located accross the Normandy region.
- FPE database: this network is held by the Fédération de Pêche de l'Eure (FPE, fishing federation). 32 time series could be collected from Hoboware sensors with an accuracy of $0.2\,°C$. These time series are located within a single province in Normandy (north-west France).
- OFB database: this network is held by the Office Français de la Biodiversité (OFB). In this study, we collected 956 time series measured by Hoboware sensors with an accuracy of $0.2\,°C$. These time series cover the whole territory.

The time series collected for this study range between 2006 and 2023. A common frequency (hourly) was chosen to combine the three datasets (Fig.2, the white color indicates missing data periods). The hourly timescale is on the y axis and the origin of the database on the x axis, so that each pixel column corresponds to one of the 993 measurement stations. On the left hand side, stations from the DREAL database, then stations from the FPE data base and on the right hand side, stations from the OFB database. The colour scale is proportional to the value of T_W. Reddish colours indicate warmer temperatures (occurring in summer) and blueish colours indicate colder temperatures (occurring mostly in winter). From this perspective, we can already see several particularities.

- Many time series are incomplete or cover short periods (less than 3 years). Therefore, a particular attention was put on keeping the treatments applicable regardless of the length of the time series.
- Several time series cover a wide range of years but with long missing periods (up to several months), especially in time series from the OFB database.
- On the contrary, several time series are relatively short (less than a year or less than 3 years) when measuring stations were abandoned shortly after their installation.
- The thermal regime is very different from one station to another. Some time series show a wide amplitude between summer and winter (displaying a high colouring contrast) whereas others experience little change over the seasons (displaying a more homogeneous colouring range).

Overall, the distribution of the stations and the characteristics of their time series provide a wide diversity of situations. The identification method developed in this study aims to address the most cases of buffered signals found in this dataset. It is also adapted to irregular data acquisition[27].

3 | THEORY

The method developed in this paper aims to 1. identify particular thermal regime such as buffered signals in time series, then 2. sort occasional and regular buffered data. A buffered signal is defined by a lower amplitude range compared to the rest of the time series. It can be caused by natural phenomena or by particular events. Identifying buffered signals in a time series can help define "normal" thermal regimes within the rest of the time series, a group of time series or a geographical area.

3.0.1 | Identification of buffered data

Buffered signals can be caused by several factors.

- Groundwater influences: some rivers are strongly affected by groundwater inflows. Water coming from these inflows experiences much smaller thermal variations along the year[28,29,30]. If the sensor is placed close to one of these inflows or if an inflow appears during the measured period, a thermal regime change can be detected[31,29].

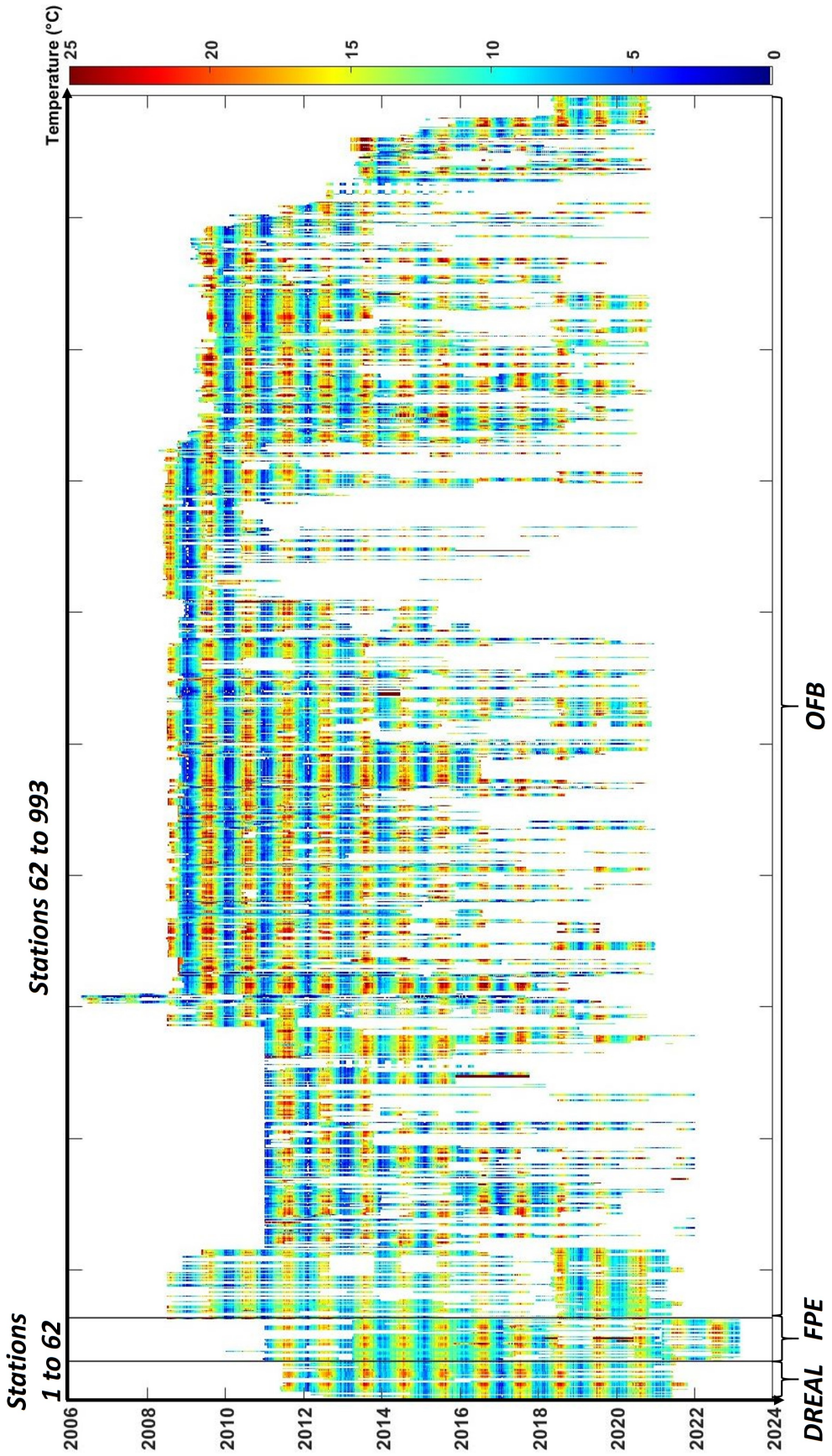


FIGURE 2 2D chart illustrating all the water temperature time series used in this study. Each column along the x-axis corresponds to one measurement station and the y-axis shows the hourly timeline. White areas indicate no data. The colourbar is proportional to the water temperature range over the studied period.

- Human facilities: during dams' operations or important floods, sediment movements can cover a sensor with a thick layer[18,14,17]. In this case, the river thermal regime does not follow the climate regime influence anymore and the amplitude strongly decreases. This event is usually seen as a sudden change of amplitude. One way to recognize this phenomenon is to compare periods when the sensor measures the water temperature again, before and after the sudden change of amplitude.
- Sensors dysfunctions: time series coming from not qualified databases can present important deficiencies such as missing periods or irrelevant data caused by human interventions. For example a sensor moved in another location or stored in a basement would measure unusual temperatures which do not properly represent a river's thermal regime. The dataset used in this study contains numerous time series with interrupted measurements. Database managers do not always give the necessary metadata to follow these interruptions, leaving such identifications steps to pre-processing operations.

Buffered data can be characterised by several aspects: a change in amplitude, a more important lag time or a different ratio to air temperature. In this paper, we define a normalized amplitude index which detects this kind of signal using only the amplitude of the time series itself: the Diurnal Temperature Range to Peak (DTRP).

$$DTRP = \frac{|T_{Wmax} - T_{Wmin}|}{T_{Wmax}} \tag{1}$$

With T_{Wmax}, the daily maximum temperature and T_{Wmin}, the daily minimum temperature. The DTRP indicates the variability of the temperature in the day normalized by the maximum temperature of the day. The normalization allows the metric to be applied on periods with different amplitude ranges and time series with different mean temperatures. The closer to 0, the less contrast between the maximum and minimum temperatures of the day. This metric requires high frequency measurements, at least on several days, but does not require lag time calculation or additional variable inputs. The identification of buffered is defined with a threshold level tol_c (Eq.2). In this study, the threshold level was set to 0.05, which means that a diurnal variation of less than 5% of the maximum temperature is considered as a buffered signal (for a daily maximum temperature of 20 °C, buffered data would be defined by a diurnal variation of less than 1 °C).

$$DTRP < tol_c \tag{2}$$

tol_c selects the level of DTRP considered as buffered data. In a time series with a naturally low amplitude, the detection limit of a buffered signal would concern most of the signal. If a time series is marked by an occasional buffered signal, the DTRP level would highlight the concerned period. This value was determined from filtering rules used on water temperature studies in metropolitan France[32] and also known time series produced in Normandy where such particularities was recently diagnosed[17]. When applied to other variables, this threshold level would be adjusted according to the amplitude dynamic of the variable.

3.1 | Occasional or regular buffered signal

Once buffered data is identified in time series, the method determines if it occurs occasionally or regularly (most of the time on a cyclical-annual basis). In the former case, further preprocessing steps can be applied to remove or replace the data if necessary. To distinguish regular buffered signals from occasional ones, the cumulative count (C) of buffered data along the time series was performed. For each time series, the number of buffered data was normalized with a cumulative density function (CDF) and then partitioned into bins. The width of the bins (w) was defined using a Freedman-Diaconis rule (Eq: 3) to be less sensitive to outliers. This method is also more suitable for non Gaussian distribution.

$$w = 2 \times IQR(T_W) \times N^{-1/3} \tag{3}$$

With IQR the interquartile range applied to the time series and N the number of elements. The cumulative count was performed using the "histcounts" function from Matlab R2022b.
For each count vector C, the standard deviation (σ_C) is then calculated. Highest values of σ_C indicate occasional buffered signals whereas middle values of σ_C indicate time series with regular buffered signal periods. Lowest values of σ_C indicate time series with little or no buffered signal. The limits proposed are set to defined highest and lowest σ_C (Eq. 4) from the mean standard

deviation of the datasets used in this study and time series with known thermal regimes as references.

$$C_{lim} = \bar{C} \pm 2 \times \sigma_C \tag{4}$$

With $\bar{C}$ the mean value of C and σ_C the standard deviation of C.

LISTING 1 Matlab command for the identification of particular buffered signal events

```
idx=find(Sing_Matrix==-1); %select buffered signal index
idxnorm=(idx-idx(1))/(idx(end)-idx(1)); %bring index values in [0 1] range
[C,edge]=histcounts(idxnorm,'Normalization','cdf','BinMethod','fd');
Sigma_C=std(C); %calculate standard deviation of cdf results
Sigma_C_cat=zeros(size(Sigma_C));
% Set -1 for time series with little buffered signal
Sigma_C_cat(Sigma_C<(mean(Sigma_C,'omitnan')-2*std(Sigma_C,'omitnan')))=-1;
% Set 1 for time series with occasional buffered signal
Sigma_C_cat(Sigma_C>(mean(Sigma_C,'omitnan')+2*std(Sigma_C,'omitnan')))=1;
```

3.2 | Seasonal analysis of regular buffered data

High amount of regular buffered signal are not necessarily linked to identical environmental factors. In this section, a classification is proposed to distinguish several types of buffered signals according to their frequency and the season of occurrence. Indeed, regular buffered often occurs on a cyclical basis related to french climates seasonal characteristics.

A histogram count is performed on the months of the time of buffered signal. Then, the major season concerning the 3 months with the highest count is extracted. As the studied area is located in the Northern hemisphere, seasons are defined as such.

- Winter: December - January - February
- Spring: March - April - May
- Summer: June - July - August
- Fall: September - October - November

For each time series concerned by a high amount of regular buffered signal, the season with the most buffered signal is thus defined. Time series are then classified according the their major buffered signal season.

LISTING 2 Matlab command for the seasonal classification of time series.

```
for j=1:NS %NS=number of stations
    if Sigma_C_cat(j)==1
        season_low(j,1)=0; %set 0 if no buffered signal
    else
        time_low=TimeFrance(Sing_Matrix_France(:,j)==-1); %extract buffered signal time
        month_low=month(time_low); %extract month for each date
        [N,edge]=histcounts(month_low,'Normalization','count','BinMethod','integers');
        [~,idx]=maxk(N,3); %select the 3 first maxima
        [~,idxm]=maxk(N,1); %select the first maximum
        month_low(j,1)=idxm; %select the month with the most buffered signal
        idxmax(j,1:length(idx))=idx;
        winter=ismember(idx,[1 2 12]);
        spring=ismember(idx,[3 4 5]);
        summer=ismember(idx,[6 7 8]);
        fall=ismember(idx,[9 10 11]);
        season=[sum(winter) sum(spring) sum(summer) sum(fall)];
```

```
        [~,idxseason]=max(season); %select the season with the most buffered signal
        season_low(j,1)=idxseason;
    end
end
```

4 | RESULTS AND APPLICATIONS

4.1 | Application of the method on two examples

As an illustration, the identification method is applied on two time series: one recorded on the *Odon* river and one recorded on the *Lot* river (Fig.1).

The first time series recorded on the *Odon* river, station n°9 is a case of a sensor irregularity caused by a sediment movement. This water temperature time series is marked by a long period of buffered data in 2017 (Fig.3a.1 grey curve). For this station, the buffered signal was interpreted, using Independant Component Analysis [17], to be caused by deposits and erosion. During the beginning of the 2017 spring, the low streamflow settled the suspended particles which progressively covered the sensor, causing an important change in the signal. Later precipitation events in 2017-2018 winter removed the sediment and the signal went back to its previous dynamics. The identification method was able to highlight the period with a perturbed signal (dark blue dots). The corresponding DTRP (solid orange line) during this period remains under the 0.05 limit (dashed orange line). In total, buffered data represents 2.82% of the recorded time series.

As a second illustration, the identification method is applied on a station characterized by regular buffer signal, especially during the second part of the year (Fig.3a.2). It is interesting to note that although the annual temperature range on the *Lot* is wider than on the *Odon*, the thermal regime is characterized by a much lower DTRP.

With the buffered data normalized cumulative count, the method identified the *Odon* time series' buffered data as occasional (Fig.3b.1) and the *Lot* time series' buffered data as regular (Fig.3b.2).

This method in two steps highlighted the buffered data in two different time series recorded in very different climate zones, different rivers and different thermal regimes then identified if it was caused by a particular event or part of the river's thermal regime.

4.2 | Application of the method at national scale

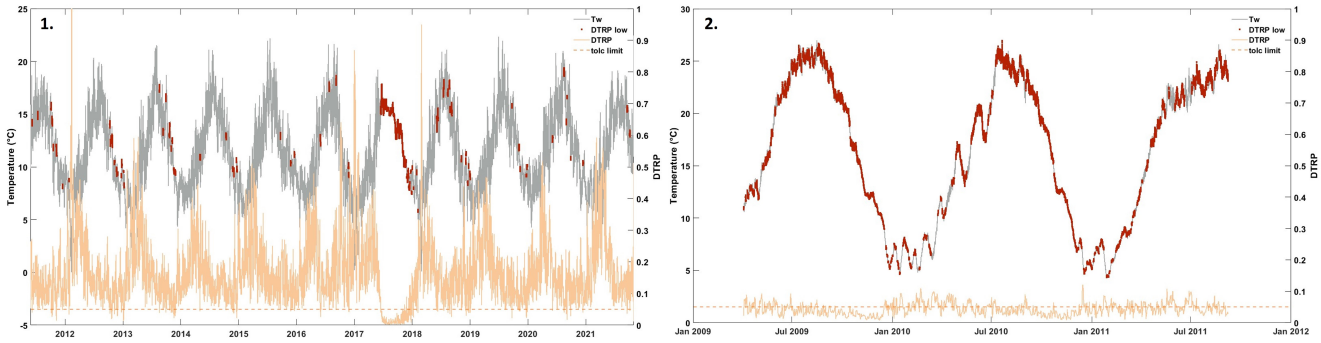
4.2.1 | Identification of buffered signal in time series

The first step of the method is the identification of buffered signals in time series. The percentage of buffered data for each of the 993 stations is presented at national scale according to the elevation (Fig.4). The percentage of buffered data was classified into 7 categories (red scale) to represent the distribution on the whole dataset. From a qualitative point of view, the North and West parts of the territory seem marked by a higher number of buffered time series. Two zones in particular seem to concentrate buffered time series: the *Seine* catchment (Northern black lined catchment) and the *Garonne* catchment (Southern black lined catchment). On the contrary, mountains areas (in red and characterized by Type 1 climate) show less buffered data. Similarly, the Atlantic facade (in blue and characterized by Type 5 climate) shows less buffered data.

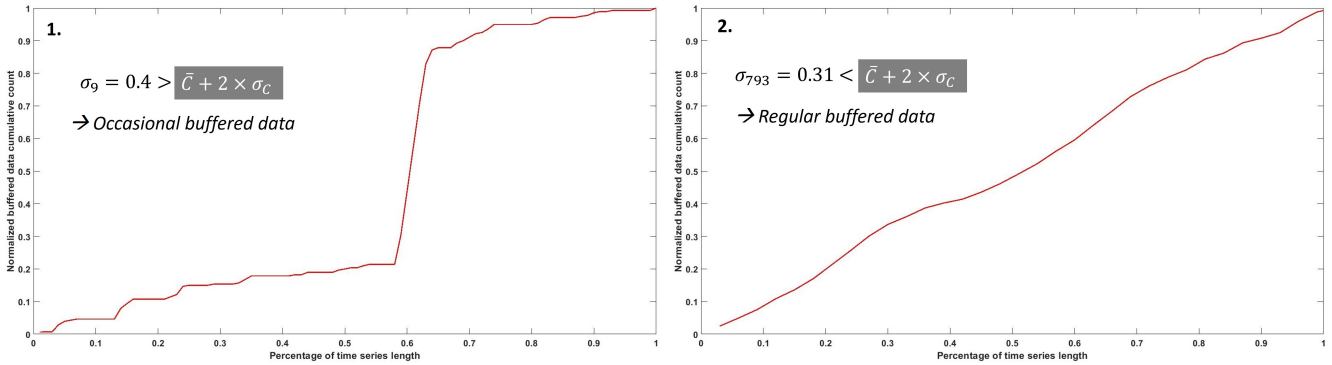
The distribution of the amount of buffered data with the climate Types was analyzed (Fig.5). A two-sample t-test confirmed the significant difference of the amount of buffered data between Types 1, 2 and 3. A significant difference was also found between Types 4, 5 and 6.

Mountain related climates (Types 1, 2, 6) include stations with less buffered data and a relatively low standard deviation for the amount of buffered data. This can be explained by lower maximum daily T_W in high altitude zones. Therefore, the 5% DTRP threshold implies a lower daily amplitude variation (for a maximum daily amplitude of $10\,^{\circ}$C, a DTRP of 5% means a daily variation of less than $0.5\,^{\circ}$C). This threshold is more difficult to reach and leads to a low amount of buffered data.

Intermediate and low altitude regions characterized by climate types 3, 4 and 7 include stations with a higher amount of buffered data and a higher standard deviation on this quantity. The increasing annual temperature increases the probability of buffered data as the 5% threshold allows for a wider daily thermal amplitude. Type 4 is characterized by the minimum annual



(a) 1. Station n°9 on the Odon river (Normandy, north west France) as an example of the identification method for buffered signal. The water temperature (grey curve, left axis) shows a period with a unusual buffered signal (blue dots) in 2017. The corresponding DTRP (solid orange line, right axis) decreases under the threshold (dashed orange line) during the buffered signal period. 2. Station n°793 on the Lot river (Nouvelle Aquitaine, south west France) as an example of the identification method for a regular buffered signal. The water temperature (grey curve, left axis) shows that most of the signal is marked by buffered signal (blue dots). The corresponding DTRP (solid orange line, right axis) regularly decreases under the threshold (dashed orange line) especially during fall.



(b) 1. Corresponding buffered data normalized cumulative count for the Odon time series. The count standard deviation above the mean standard deviation of the dataset identifies the buffered data as occasional. 2. Corresponding buffered data normalized cumulative count for the Lot time series. The lower count standard deviation identifies a time series with regular buffered data.

FIGURE 3 Illustration of the identification method with two time series. (a) Recorded time series with buffered data identified in blue. Corresponding DTRP in orange. (b) Cumulative count of buffered data to identify regular from occasional buffered data. 1. A time series recorded on the Odon river with an occasional episode of buffered data. 2. A time series recorded on the Lot river with regular buffered data.

amplitude and a high mean air temperature which favors the occurrence of buffered data in T_W. The higher standard deviation in the amount of buffered data can be explained by a higher inter-annual variability in these 3 Types (3, 4 and 7) compared to mountain related climates. However, this variability is to be compared in the coming decades with the growing influence of climate change in mountain areas.

The Atlantic facade and oceanic climate (Type 5) contains mostly stations with a low amount of buffered data despite similar altitude with Type 4. Type 5 is characterized by a low inter-annual temperature variability (thus leading to a reduced standard deviation for the amount of buffered data). However, a more important precipitation rate adds disturbing factors in the T_W daily amplitude.

Type 8 is very contrasted, both in the thermal and precipitation regimes, with high inter-annual variability for both of them. This leads to a high standard deviation of the amount of buffered data.

The distribution of buffered data on a national scale can be partly explained by climate types. However, buffered data can also be caused by factors independent from climate variables such as local characteristics (elevation, groundwater inflows, riparian

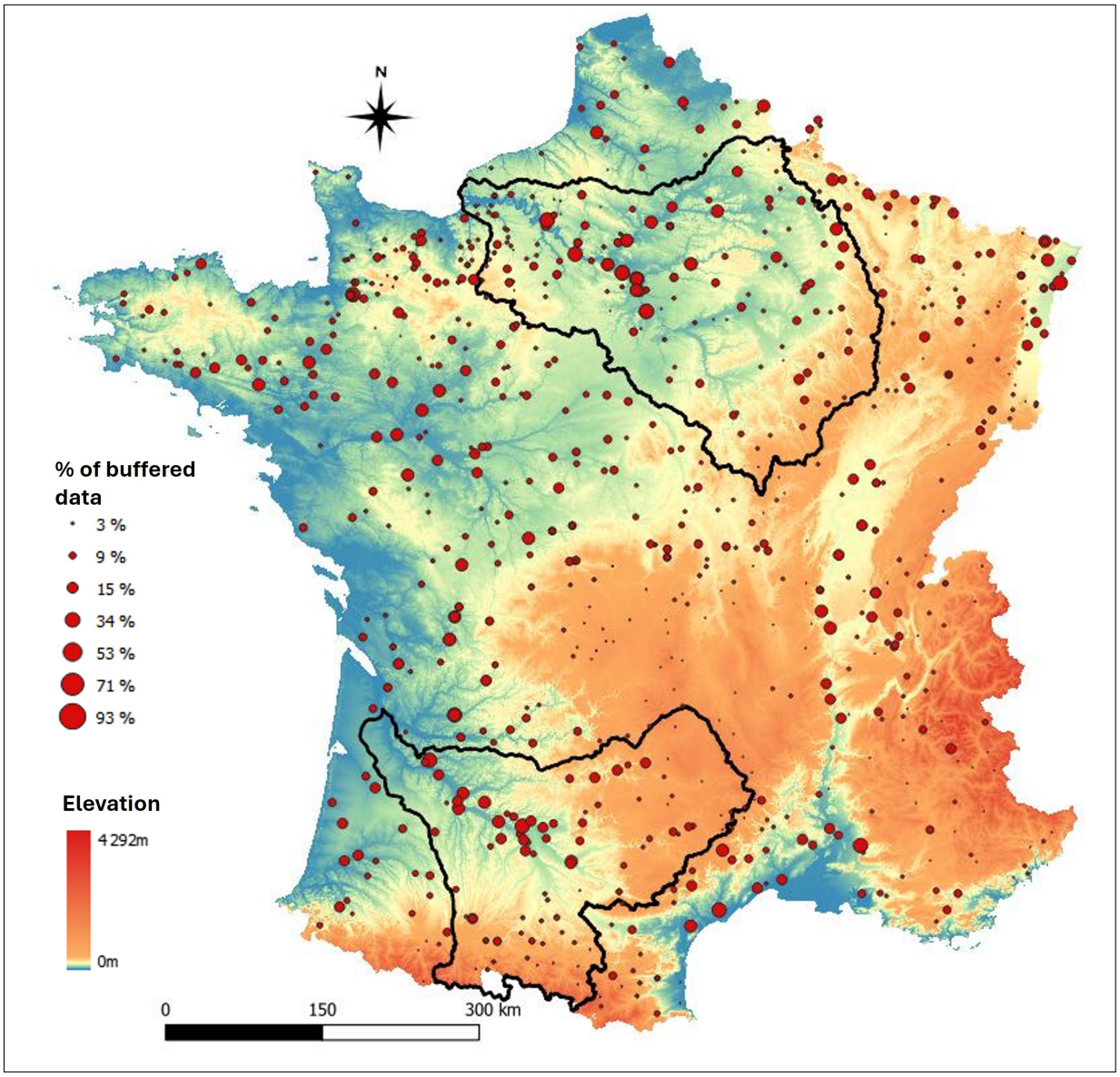


FIGURE 4 Percentage of buffered data for each time series of the dataset. The *Seine* catchment (North) and *Garonne* (South) catchment gather stations with the most buffered data. High altitude regions (red) contain stations with the least amount of buffered data.

vegetation or the presence of urban areas). Examples on 4 major rivers show that low altitudes stations tend to be characterized by more buffered data (Fig.6). Furthermore, in this study, a distinction is made between occasional buffered data and regular buffered data. Climate types are more prone to explain the latter category.

4.2.2 | Sorting stations with occasional buffered data

The second step of the method aims to sort time series with occasional buffered data from time series with regular buffered data. As described in the theory section, the normalized count over buffered data days is performed on the whole dataset and the standard deviation of the count is taken (Fig.7). As shown by the colour pattern, all climate types can be concerned by occasional as well as regular buffered data. However, with the use of limits, time series with an extreme standard deviation can be discriminated.

For this study, the limits were set to sort the 5% extreme quantiles of the count standard deviation distribution. These limits

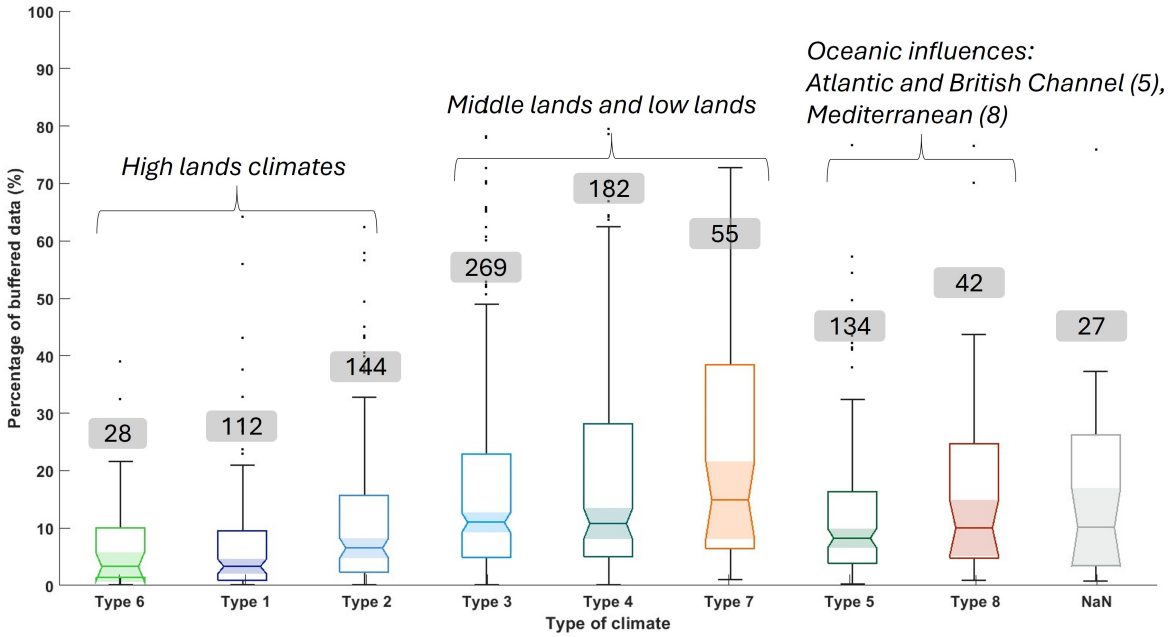


FIGURE 5 Percentage of buffered data for each climate type (y axis). Squared figures indicate the number of stations in each climate category.

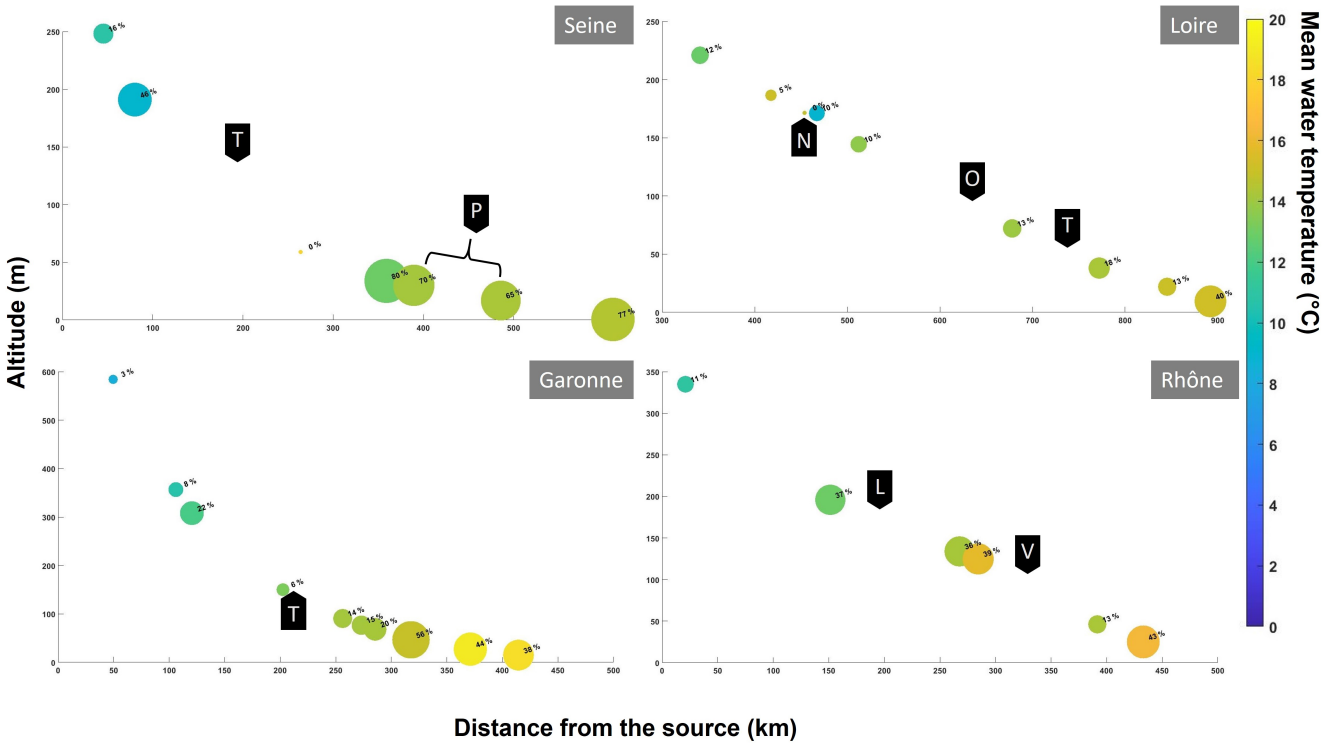


FIGURE 6 Mean water temperature (colorscale) and percentage of buffered data (dot area scale) along 4 major rivers in France. Black labels indicate main cities crossed by the river. *Seine*: T=Troyes, P=Paris. *Loire*: N=Nevers, O=Orléans, T=Tours. *Garonne*: T=Toulouse. *Rhône*: L=Lyon, V=Valence.

allowed to correctly categorize time series with known occasional buffered data periods. In total, 41 time series were identified with little or no buffered data and 12 were identified with occasional buffered data.

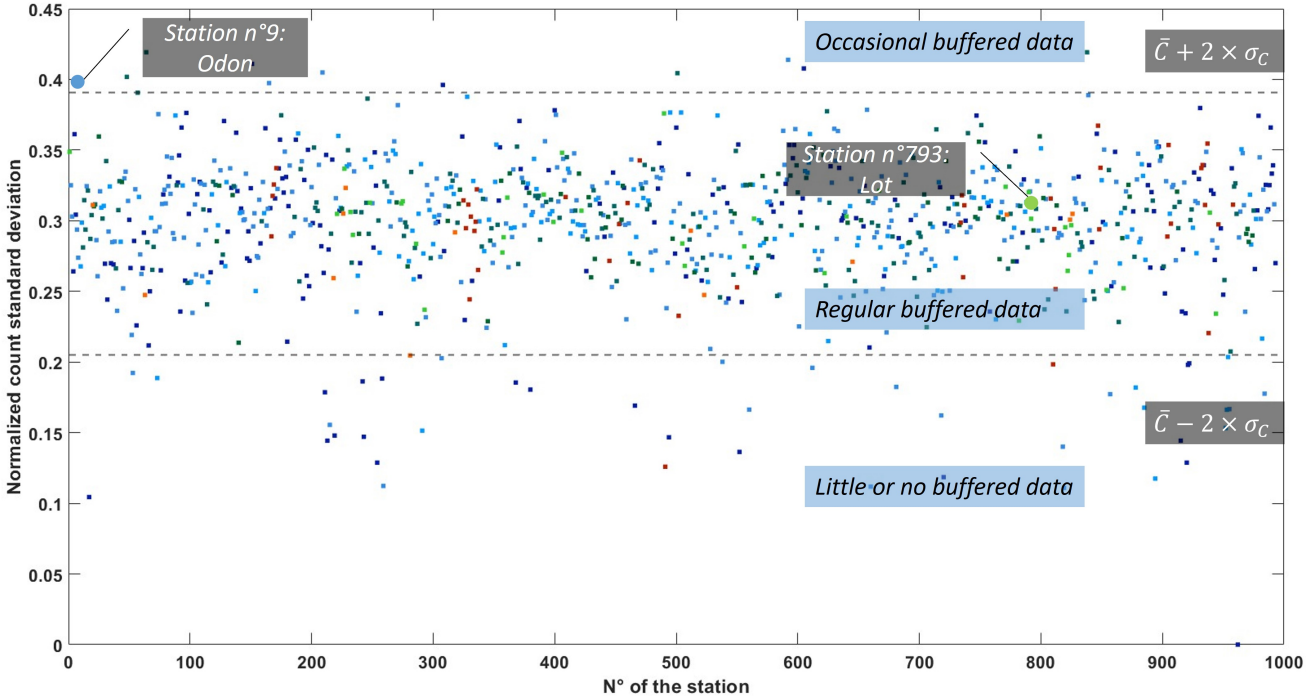


FIGURE 7 Standard deviation of the normalized count of the buffered days. The colour scale indicate the type of climate. Dashed lines show the upper limit (which indicates time series with occasional buffered data) and lower limit (which indicates time series with little or no buffered data). The middle part gathers time series with buffered data categorized as regular. The two labelled stations are illustrative cases mentioned earlier.

4.3 | Monthly and seasonal distribution of the regular buffered signals

In this section, only time series with regular buffered data (stations located in the middle section in Fig.7) are considered. In each time series, for each day characterized by buffered data, months are extracted. The main buffered season is determined among the three months with the most buffered days. In metropolitan France, with this dataset, most stations experience buffered data in fall, especially in November (Fig.8) followed by winter, especially in January. In October and November, stored heat contained in the soils can buffer the thermal amplitude generated by the day/night air temperature cycle. During winter, the soil influence is less present so rivers' thermal regime is more sensitive to atmospheric factors. Therefore, the amount of buffered data depends on the thermal amplitude for each climate. In February, however, the daily thermal amplitude starts to increase and reduces the probability of buffered data.

Type 1 is the only climate type with more buffered data in winter than in fall. Then Type 2 counts a Winter/Fall ratio of 0.6. Type 8 and 5 count a ratio around 0.4. Types 3, 4 and 6 count a ratio of 0.3. The climate type with the most determined buffered season is Type 7 with a Winter/Fall ratio of 0.08, which means that most of its buffered data occurs in fall in this region.

Spring is the major buffered season for the least number of stations. This can be explained by a higher thermal amplitude between day and night during this season in metropolitan France. Furthermore, soils have not yet stored heat to buffer this thermal amplitude (as in Fall for example). However, Type 2 seems the most concerned by buffered data in March. This climate type is located at the foot of mountain area. Rivers' thermal regime there can be subject to snow melt influences during this period.

The spatial distribution mostly highlights that the *Seine* catchment counts the greatest diversity of buffered seasons (Fig. 9). This diversity can partly be explained by the diversity of climate types in this catchment (Types 1,2,3,4,5).

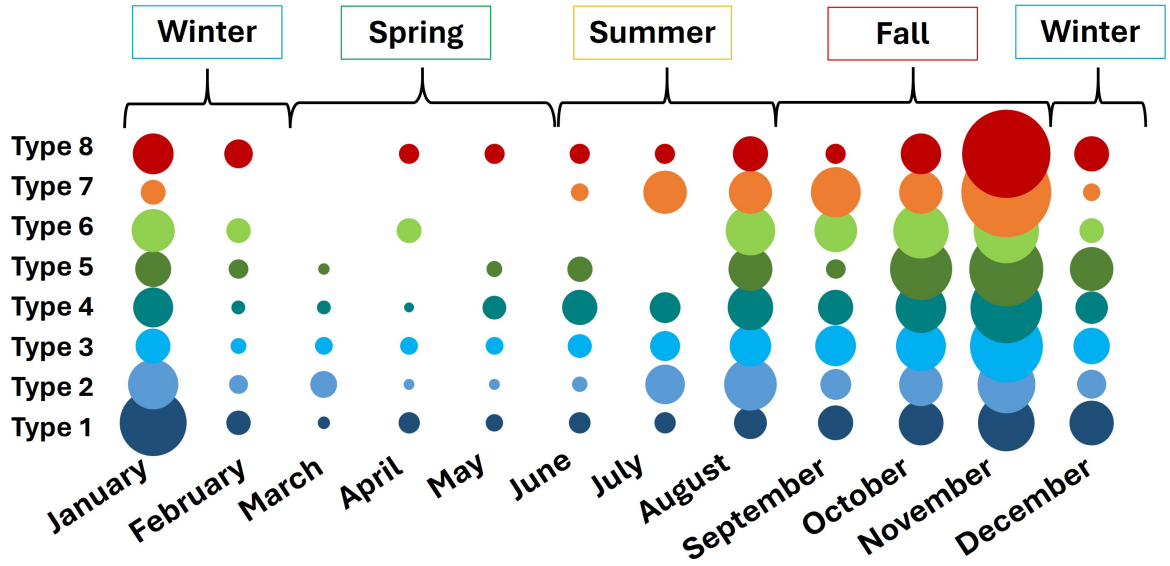


FIGURE 8 Determination of months most concerned with buffered data for each climate type defined in[26]. The size of each circle is proportional to the percentage of stations in each climate type on the y axis concerned by the most buffered month on the x axis.

Understanding streams and rivers' temporal evolution is a necessity to protect their vulnerabilities[33,19]. In this regard, time series are powerful tools to provide information. Inside time series, singular behaviours often occur and can represent either obstacles to understand a river's thermal regime or relevant information[21,2].
In this paper, an identification method based on a normalized diurnal amplitude range is proposed to highlight singularities characterized by reduced amplitude variations compared to a normal thermal regime. Normal thermal regimes are defined by typical non buffered periods in the time series itself or by non buffered time series measured in the same geographical context. These singularities are less described than outliers in the literature and can correspond to various phenomena. Contrary to outliers, buffered signals can contain valuable information. That is why, the identification method was applied at national scale on a dataset of 993 water temperature measurement stations.
Natural buffering phenomena and particular sensor behaviours can display similar signals. However, a major difference between these behaviours lies in the temporal scale. Sensor particular behaviours are expected to occur on shorter periods not related to the river thermal regime. Natural buffering phenomena on the contrary are more prone to appear regularly over a hydrological year or be related to other hydrological cycles (such as the precipitation regime, streamflow, conductivity etc). The method proposed in this study distinguishes occasional buffered signal from regular ones.

Despite the large amount of data used to develop this identification method, the qualification of "buffered data" relies on the threshold parameter tol_c. The value proposed in this paper is adapted to metropolitan France climatic zones. In a country with totally different climate types, this value may need to be adjusted. A solution could be to define this threshold parameter prior to the preprocessing using climate and geographical data.
Improvements would include the application of the method in other countries with different climates to measure the variability of the threshold parameter. It would also allow to find a relationship between the calibrating parameter and climate variables and therefore, predict its value for a new country with fewer monitoring. Also, applying this method to datasets with other types of variables (streamflow, conductivity, precipitation) could extend the understanding of mechanisms leading to buffered signals.

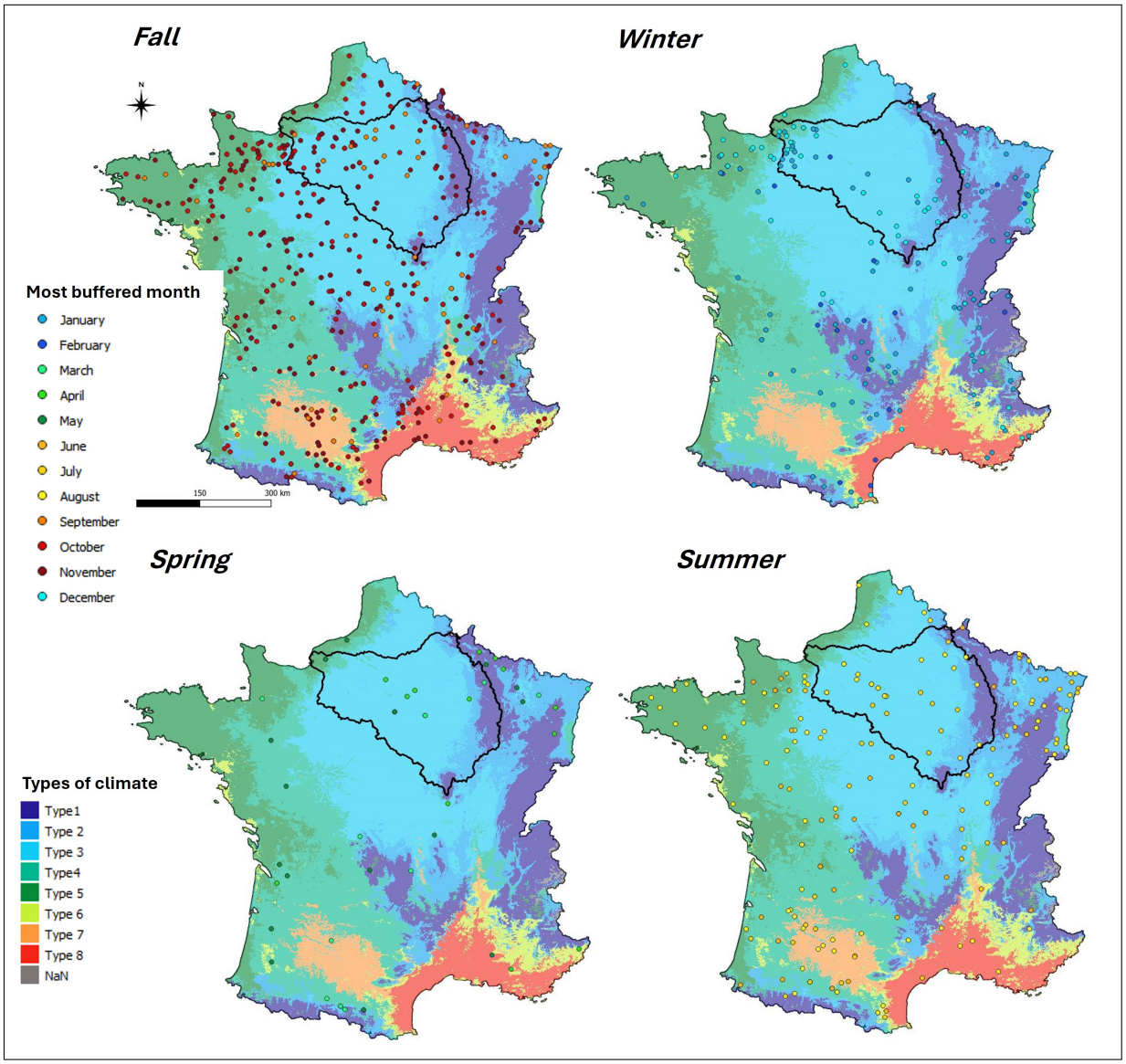


FIGURE 9 Distribution of stations characterized by the major buffered month categorized by seasons according to climate types. The *Seine* catchment (black) gathers the greatest diversity of buffered seasons.

5 | CONCLUSION

When monitoring environmental parameters (temperature, pH, chemical components...), numerous sensors are often needed in order to get accurate spatial and temporal information. However, the measured data sometimes come along with several defects or unusual signals. Some are caused by sensors' malfunctions whereas some may convey useful information. Preprocessing data can lead to the loss of information and misinterpretations of the data. In this paper, an identification method is proposed as a first step before preprocessing data. This identification method sorts out buffered signal defined as unusual lower amplitude signals compared to the rest of the time series. In this study, the method was applied at national scale on a dataset with a variety of data quality in order to cover 1. numerous cases of data failures, 2. time series recorded in a variety of climate types and 3. a diversity of river environments.

The identification method highlights particular periods or time series due to measurement anomalies or caused by natural phenomena. It uses operations which need only the time series itself and can be applied without using additional data.

Identify singularities in time series before filtering them allows to better understand singular behaviours in a measurement

station. It also provides elements of interpretations to understand a river's thermal regime. When integrating time series in modelling operations, identify the cause of buffered signals can give insights on environmental influencing factors. Applied on a larger scale (several climate countries for example), this method can compare thermal behaviours trends between regions and/or give information on the quality of the data between different databases. Buffered signals are not exclusive to water temperature time series. The identification of this behaviour in other variables (air temperature, diurnal heat transfer, carbon emission from biomass...) could highlight other types of regime changes.

AUTHOR CONTRIBUTIONS

Nelly Moulin: Conceptualization, Methodology, Writing. Frederic Gresselin: Conceptualization, Resources, Review and Editing. Bruno Dardaillon: Conceptualization, Validation. Zahra Thomas: Conceptualization, Funding Acquisition, Resources, Review and Editing.

ACKNOWLEDGMENTS

The authors thank the Direction Régionale de l'Environnement, de l'Aménagement et du Logement (DREAL) de Normandie as well as the Office Français de la Biodiversité and the Fédératon de Pêche de l'Eure for the availability of the measurement data.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Cortés-Ibáñez JA, González S, Valle-Alonso JJ, Luengo J, García S, Herrera F. Preprocessing methodology for time series: An industrial world application case study. *Information Sciences.* 2020;514:385–401. doi: 10.1016/j.ins.2019.11.027
2. Di Blasi JIP, Martínez Torres J, García Nieto PJ, Alonso Fernández JR, Díaz Muñiz C, Taboada J. Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NW Spain). *Ecological Engineering.* 2013;60:60–66. doi: 10.1016/j.ecoleng.2013.07.054
3. Daufresne M, Boët P. Climate change impacts on structure and diversity of fish communities in rivers. *Global Change Biology.* 2007;12(13):2467–2478. Publisher: Wiley.
4. Whitehead PG, Wilby RL, Battarbee RW, Kernan M, Wade AJ. A review of the potential impacts of climate change on surface water quality. *Hydrological Sciences Journal.* 2009;54(1):101–123. doi: 10.1623/hysj.54.1.101
5. Comte L, Buisson L, Daufresne M, Grenouillet G. Climate-induced changes in the distribution of freshwater fish: observed and predicted trends: *Climate change and freshwater fish*. *Freshwater Biology.* 2013;58(4):625–639. doi: 10.1111/fwb.12081
6. Hannah DM, Garner G. River water temperature in the United Kingdom: Changes over the 20th century and possible changes over the 21st century. *Progress in Physical Geography: Earth and Environment.* 2015;39(1):68–92. doi: 10.1177/0309133314550669
7. Magnuson JJ, Crowder LB, Medvick PA. Temperature as an Ecological Resource. *American Zoologist.* 1979;19(1):331–343. doi: 10.1093/icb/19.1.331
8. Caissie D. The thermal regime of rivers: a review. *Freshwater Biology.* 2006;51(8):1389–1406. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2427.2006.01597.xdoi: 10.1111/j.1365-2427.2006.01597.x
9. Dugdale SJ, Malcolm IA, Kantola K, Hannah DM. Stream temperature under contrasting riparian forest cover: Understanding thermal dynamics and heat exchange processes. *Science of The Total Environment.* 2018;610-611:1375–1389. doi: 10.1016/j.scitotenv.2017.08.198
10. Beaufort A, Moatar F, Sauquet E, Loicq P, Hannah DM. Influence of landscape and hydrological factors on stream–air temperature relationships at regional scale. *Hydrological Processes.* 2020;34(3):583–597. doi: 10.1002/hyp.13608
11. Heddam S. Chapter 13 - Outlier robust extreme learning machine: Predicting river water temperature in the absence of air temperature. In: Eslamian S, Eslamian F., eds. *Handbook of Hydroinformatics*, , Elsevier, 2023:205–221
12. Gong W, Yang D, Gupta HV, Nearing G. Estimating information entropy for hydrological data: One-dimensional case. *Water Resources Research.* 2014;50(6):5003–5018. doi: 10.1002/2014WR015874
13. Arismendi I, Safeeq M, Johnson SL, Dunham JB, Haggerty R. Increasing synchrony of high temperature and low flow in western North American streams: double trouble for coldwater biota?. *Hydrobiologia.* 2013;712(1):61–70. doi: 10.1007/s10750-012-1327-2
14. Chandesris A, Van Looy K, Diamond JS, Souchon Y. Small dams alter thermal regimes of downstream water. *Hydrology and Earth System Sciences.* 2019;23(11):4509–4525. doi: 10.5194/hess-23-4509-2019
15. Poirel J, Gailhard J, Capra H. Influence des barrages-réservoirs sur la température de l'eau : exemple d'application au bassin versant de l'Ain. *La Houille Blanche - Revue internationale de l'eau.* 2010;4:p. 72 – p. 79. Publisher: EDP Sciencesdoi: 10.1051/lhb/2010044
16. Seyedhashemi H, Moatar F, Vidal JP, et al. Thermal signatures identify the influence of dams and ponds on stream temperature at the regional scale. *Science of The Total Environment.* 2021;766:142667. doi: 10.1016/j.scitotenv.2020.142667
17. Moulin N, Gresselin F, Dardaillon B, Thomas Z. River temperature analysis with a new way of using Independant Component Analysis. *Frontiers in Earth Science.* 2022;10.
18. Kędra M, Wiejaczka . Climatic and dam-induced impacts on river water temperature: Assessment and management implications. *Science of The Total Environment.* 2018;626:1474–1483. doi: 10.1016/j.scitotenv.2017.10.044

19. Ouellet V, St-Hilaire A, Dugdale SJ, Hannah DM, Krause S, Proulx-Ouellet S. River temperature research and practice: Recent challenges and emerging opportunities for managing thermal habitat conditions in stream ecosystems. *Science of The Total Environment.* 2020;736:139679. doi: 10.1016/j.scitotenv.2020.139679
20. Moatar F, Gailhard J. Water temperature behaviour in the River Loire since 1976 and 1881. *Comptes Rendus Geoscience.* 2006;338(5):319–328. doi: 10.1016/j.crte.2006.02.011
21. Cho HY, Oh JH, Kim KO, Shim JS. Outlier detection and missing data filling methods for coastal water temperature data. *Journal of Coastal Research.* 2013;165:1898–1903. doi: 10.2112/SI65-321.1
22. Hasan EA. A Method for Detection of Outliers in Time Series Data. *International Journal of Chemistry, Mathematics and Physics.* 2019;3(3):56–66. doi: 10.22161/ijcmp.3.3.2
23. Narajewski M, Kley-Holsteg J, Ziel F. tsrobprep — an R package for robust preprocessing of time series data. *SoftwareX.* 2021;16:100809. doi: 10.1016/j.softx.2021.100809
24. Höppner F. Improving time series similarity measures by integrating preprocessing steps. *Data Mining and Knowledge Discovery.* 2017;31(3):851–878. doi: 10.1007/s10618-016-0490-x
25. Marti-Puig P, Blanco-M A, Cárdenas JJ, Cusidó J, Solé-Casals J. Effects of the pre-processing algorithms in fault diagnosis of wind turbines. *Environmental Modelling & Software.* 2018;110:119–128. doi: 10.1016/j.envsoft.2018.05.002
26. Joly D, Brossard T, Cardot H, Cavailhes J, Hilal M, Wavresky P. Les types de climats en France, une construction spatiale. *Cybergeo: European Journal of Geography.* 2010. Publisher: CNRS-UMR Géographie-cités 8504doi: 10.4000/cybergeo.23155
27. Ouarda TBMJ, Charron C, St-Hilaire A. Regional estimation of river water temperature at ungauged locations. *Journal of Hydrology X.* 2022;17:100133. doi: 10.1016/j.hydroa.2022.100133
28. Evans EC, McGregor GR, Petts GE. River energy budgets with special reference to river bed processes. *Hydrological Processes.* 1998;12(4):575–595. doi: 10.1002/(SICI)1099-1085(19980330)12:4<575::AID-HYP595>3.0.CO;2-Y
29. Le Lay H, Thomas Z, Rouault F, Pichelin P, Moatar F. Characterization of Diffuse Groundwater Inflows into Stream Water (Part II: Quantifying Groundwater Inflows by Coupling FO-DTS and Vertical Flow Velocities). *Water.* 2019;11(12):2430. Number: 12 Publisher: Multidisciplinary Digital Publishing Institutedoi: 10.3390/w11122430
30. Gresselin F, Dardaillon B, Bordier C, Parais F, Kauffmann F. Use of statistical methods to characterize the influence of groundwater on the thermal regime of rivers in Normandy, France: comparison between the highly permeable, chalk catchment of the Touques River and the low permeability, crystalline rock catchment of the Orne River. *Geological Society, London, Special Publications.* 2021:SP517–2020–117. doi: 10.1144/SP517-2020-117
31. Lalot E, Curie F, Wawrzyniak V, et al. Quantification of the contribution of the Beauce groundwater aquifer to the discharge of the Loire River using thermal infrared satellite imaging. *Hydrology and Earth System Sciences.* 2015;19(11):4479–4492. doi: 10.5194/hess-19-4479-2015
32. Moatar F, Sauquet E, Magand C. Thermie en rivière : Analyse géostatistique et description de régime : Application à l'échelle de la France. tech. rep., Université de Tours GéHCO; INRAE UR RiverLy: 2020.
33. Kaandorp VP, Doornenbal PJ, Kooi H, Peter Broers H, Louw dPGB. Temperature buffering by groundwater in ecologically valuable lowland streams under current and future climate conditions. *Journal of Hydrology X.* 2019;3:100031. doi: 10.1016/j.hydroa.2019.100031