Supplement of The Cryosphere, 17, 761–787, 2023
https://doi.org/10.5194/tc-17-761-2023-supplement

The Cryosphere Open Access EGU

Supplement of

Channelized, distributed, and disconnected: spatial structure and temporal evolution of the subglacial drainage under a valley glacier in the Yukon

Camilo Andrés Rada Giacaman and Christian Schoof

Correspondence to: Camilo Andrés Rada Giacaman (camilo@rada.cl)

S1 Clustering calibration, validation, and testing

To calibrate, validate and test the clustering approach, we manually identified clusters in 12-day time windows from August 2010 to August 2015. We choose here a window length longer than the 6-day interval used in our final analysis to reduce the number of windows to the point where we are able to process them manually while still covering the whole dataset available after the 2015 field campaign. Over 461 windows, we identified a total of 613 clusters. In this process, we used all the data available to assess if a group of sensors was likely to constitute a hydraulic or mechanical cluster. Therefore, if the similarity between the diurnal residuals was not clear but suggestive of a cluster, we consulted other data sources to inform the decision. Those sources included the raw pressure records and the positions of the corresponding boreholes. For example, if a group of boreholes seemed to form a hydraulic cluster but the similarities in their diurnal residuals were only marginal, we would also study the spatial distribution of the boreholes. If we found that the boreholes were very far apart across the width of the glacier, we would then exclude the cluster, as such a configuration is unlikely for boreholes interacting through hydraulical connections. Analogously, ambiguous clusters were excluded or included based also on the similarities in their raw pressure records.

It is important to note that similarity to the eye, as when performing manual cluster identification, is not the same as similarity in the strict sense defined by the Absolute Euclidean distance. Both approaches are consistent, but to the eye the similarity seems higher when the shapes of the time series shows more numerous and distinct features. As an example, imagine two pairs of time series A1, A2 and B1, B2, such that the Euclidean distance between A1 and A2 is the same as between B1 and B2. A1 and A2 are mostly flat and show only one broad peak over the entire time window. In contrast, B1 and B2, show a very distinct pattern of multiple diurnal peaks each one with a particular shape. In this example, B1 and B2 will appear to the eye to have a higher degree of similarity than A1 and A2.

After the manual clustering of this portion of the dataset, we then divided the 613 identified clusters into three subsets of approximately 204 clusters each. These were used for the calibration, validation, and testing of the clustering method. The manual identification of clusters on what constitutes almost the entirety of the South Glacier dataset was laborious, but necessary to obtain the three independent and statistically significant sets of clusters necessary for the calibration, validation, and testing. Then, after we automated the process, we applied it to the entire data set, over shorter time windows, and with a larger overlap between windows. Allowing us to get finer temporal resolution with minimal manual intervention.

We will describe now the calibration, validation, and testing of the clustering technique used to describe South Glacier borehole data in the main text. This is agglomerative hierarchical clustering (Rokach and Maimon, 2005) over diurnal residual time series (Eq. 1) using absolute Euclidean distance metric (Eq. 2) and average-link linkage (Rokach and Maimon, 2005). Later, in section S5, we describe how we choose this particular clustering strategy from all the other available options.

The calibration consists of finding the maximum absolute Euclidean distance between time series that we believe to be associated with the same drainage subsystem or mechanical cluster. This is equivalent to finding the Split Point (SP) that best reproduce our manually picked clusters. Such SP, is then referred to as the Optimal Split Point (OSP). To find the OSP, we compared the Relative Information Gain (RIG, see Section S4) achieved by a wide range of possible Split Points (SPs) .

Using the 205 clusters of the calibration subset, we found an Optimal Split Point (OSP) of 2.26 day^{-1}, which reached a RIG of 89%. The method was then applied to the 203 clusters of the validation set, achieving a RIG of 89% as well. The validation dataset serves the purpose of checking that we have not over-fitted our calibration data. However, as the RIG achieved by the validation dataset is similar to the one by the calibration dataset, no over-fitting is suspected and no modification of the clustering parameters is needed. On the remaining 205 clusters, which correspond to the testing set, the RIG achieved was of 84%.

Alternatively, we could use other clustering techniques. In addition to hierarchical clustering, we tested Empirical Orthogonal Functions (EOFs) (Jolliffe, 2002), Self-Organizing Maps (SOMs) (Vesanto et al., 2000), and K-means clustering (David and Vassilvitskii, 2007), concluding that hierarchical clustering is the most suitable for this application. More details of the advantages and limitations of each approach can be found in Section S2. To define the best strategy to use within the hierarchical clustering framework besides the diurnal residual, we also evaluated the effectiveness of clustering using the raw pressure time series, and other pre-processing such as diurnal running mean, diurnal running standard deviation and the power spectrum. The following list describes each of the evaluated pre-processings procedures:

- **Raw** The pressure is directly fed to the clustering algorithm.
- **Diurnal running mean** The pressure signal is smoothed using a running mean with a 24 hours window.
- **Diurnal residual** The residual is computed by subtracting the diurnal running mean from the raw pressure data. A normalized version of the diurnal residual is presented in Eq. 1.
- **Residual standard deviation** A 24 hours running standard deviation (STD) computed over the diurnal residual.
- **Power spectrum** Power spectrum of the raw pressure in the frequency space.

Each of these pre-processing can additionally be "standardized", that corresponds to the normalization to the standard deviation

As an alternative distance metric, we also evaluated the performance of Euclidean distance, Correlation, and Dynamic Time Wrapping (DTW) (Mullin, 1983). DTW is a technique often used in voice recognition systems, allowing the user to measure the similarities between waveforms ignoring changes in speech speed between the reference waveform and the one being classified. In our context, DTW can help us to recognize similar records that have a small time shift (due to datalogger clock offsets), or a varying time lag due the propagation of the pressure signal in a diffusive medium. However, DTW can also erroneously match two signals with different periods if they happen to have a similar shape over the analyzed time window. A detailed description of each metric is presented in Section S3, and their performance is evaluated in Section S5.

We also evaluated two alternative linkage methods in addition to average-link. In particular, complete-link and single-link (Rokach and Maimon, 2005). These methods define the distance between two clusters respectively as the longest and shortest distance between members of each cluster. Finally, in addition to the distance Split Point criterion, we also evaluated the use of gap size and inconsistency (Zahn, 1971). Gap size SP identifies clusters in the dendrogram by finding the ones with the largest distance gap between its top node and the following node. In contrast, inconsistency identify clusters based on the "relative distance". This relative distance is obtained by normalizing the distance by the average distance of all other nodes at the same hierarchical level.

Figure S1 summarizes the performance achieved by each combination of these hierarchical clustering alternatives. We see that diurnal residual pre-processing is by far the best at reproducing hand-picked clusters. Among the distance metrics, absolute Euclidean performs better than DTW at a fraction of the computational expense. For the two leading linkage methods, the performance differences were small. However, average-link performs significantly better than complete-link when used in combination with the absolute Euclidean distance metric. Finally, the distance SP criterion performed best at identifying clusters in a dendrogram. More details about each evaluated option and the performance over the calibration and test data subsets can be found in Section S5.

S2 Alternative clustering and data analysis techniques

The hierarchical clustering technique presented in the main text was not adopted arbitrarily. On the contrary, it was the result of the careful testing and comparison of several alternative approaches. Here we will present each one of those approaches, explain why we choose hierarchical clustering and then describe in detail the different hierarchical clustering variations we tested and how we compared their performance.

The data exploration approaches presented here are not restricted only to clustering techniques, but to all the statistical methodologies we have tested in the search for tools that can help to reduce the complexity of the dataset and expose the spatial structure and phenomenon underlying the complex response observed in boreholes water pressures. The following subsections describe each evaluated methodologies.

S2.1 Empirical Orthogonal Functions (EOFs)

EOFs, also known as Principal Component Analysis (PCA) was our first attempt to reduce the complexity of the dataset and find the main modes that drive the pressure signals. After attempting the EOF analysis using the raw pressure and multiple

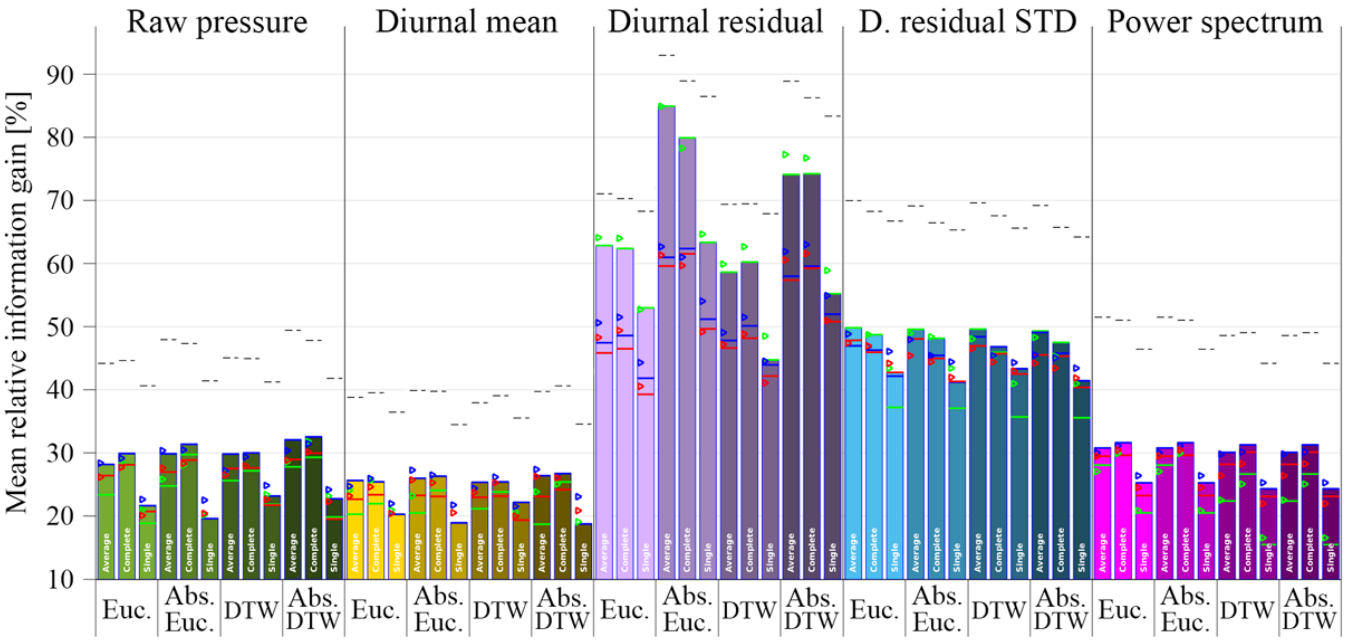

Figure S1. Mean Relative Information Gain (RIG) achieved over the 203 clusters of the validation dataset by each evaluated hierarchical clustering strategy. Coloured triangles point to the reference RIG obtained on the calibration dataset. Dashed black lines show the maximum RIG that a perfect dendrogram split method could reach on each case. The actual mean RIG achieved by each split method is shown for depth OSP (green), inconsistency OSP (Blue) and gap OSP (red).

pre-processing options such as standardized raw pressure, daily running mean, diurnal residual, and residual STD. We found that the similar sets of boreholes that we were hoping to capture in a single principal vector, usually corresponded to small principal values, because most of the variance in the dataset was explained by disconnected sensors showing large pressure variations.

After a thoughtful analysis of the results, EOFs proved of little help to understand the ongoing processes due to three main reasons:

1. The difficulty to link a specific process or signal (sometimes very well represented by one or more boreholes) with one eigenvector.

2. The tendency of single boreholes with high variability to take over the biggest eigenvalues.

3. The significant change in the eigenvectors as a result of changes in the domain (when one sensor was added or removed to the set). Making it difficult to follow one particular group of sensors in time.

EOF analysis was able to group similar pressure signals together reasonably well. However, it did not help in reducing the complexity of the data as expected. Also, EOFs did not provide a simple way to identify which principal vectors were associated to groups of very similar boreholes or groups of very distinct ones. Given that our objective is to track the evolution of these groups in time, we need a technique capable to deal with a changing number of sensors in the data set. The change of the domain size is a known problem in EOF analysis, as the principal vectors are likely to change dramatically in such circumstances, making it difficult to follow a group of similar boreholes trough time.

Despite the countless improvements that can be done to our algorithms to analyze EOF data, we decided to try more advanced techniques that overcome some of the problems presented by EOF for our particular application.

S2.2 Covariance analysis

One of the simplest approaches to identify efficient hydraulic connections (i.e. finding boreholes showing identical or very similar pressure pattern profiles), is trough their covariance. We calculated covariances matrices (equations S1 and S4) over a moving time window, including all the boreholes presenting valid data in the window interval. High covariance cases we examined to assess their actual similarity. The exercise was repeated with normalized covariances (or correlation coefficients, equations S2 and S6), which produced better results.

The covariance Cov between two time series $P^{(m)}$ and $P^{(n)}$ with N samples is given by

$$Cov(P^{(m)}, P^{(n)}) = \frac{1}{N-1} \sum_{i=0}^{N} \left(P_i^{(m)} - \overline{P^{(m)}} \right) \left(P_i^{(n)} - \overline{P^{(n)}} \right) \tag{S1}$$

where $\overline{P^{(m)}}$ is the mean value of the time series m. If $m = n$ the above equation leads to the variance of $P^{(m)}$, or $Var(P^{(m)})$.

Also, we can obtain the normalized covariance (correlation coefficient) by

$$Corr(P^{(m)}, P^{(n)}) = \frac{Cov(P^{(m)}, P^{(n)})}{Var(P^{(m)}) Var(P^{(n)})} \tag{S2}$$

We defined also the covariance matrix $\mathbf{P}$. For a set of M time series each with N samples, each column j of $\mathbf{P}$ corresponds to one time series and each row i corresponds to the samples at one time step for all time series. For simplicity we will define also the matrix $\mathbb{P}$ were the mean of each column have being removed, this is

$$\mathbb{P}_{ij} = \mathbf{P}_{ij} - \overline{P^{(j)}} \tag{S3}$$

now the covariance matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \frac{\mathbb{P}\mathbb{P}'}{N-1} \tag{S4}$$

where $\mathbf{C}_{ij} = Cov(P^{(i)}, P^{(j)})$. Let's now define the column vector D composed of the square root of the inverse of the diagonal elements of $\mathbb{P}\mathbb{P}'$, this is

$$D_i = \sqrt{\frac{1}{(\mathbb{P}\mathbb{P}')_{ii}}} \tag{S5}$$

Then, the normalized covariance matrix $\mathbb{C}$ would be given by

$$\mathbb{C} = \mathbb{P}\mathbb{P}' \odot DD' \tag{S6}$$

where $\odot$ represent the Hadamard element-wise product of matrices. Then we have $\mathbb{C}_{ij} = Corr(P^{(i)}, P^{(j)})$.

We developed a special visualization tool of the spatial distribution of the magnitude of the covariance, which allowed us to identify multiple sets of boreholes showing very similar pressure time series, suggesting they are connected or responding to the same forcing. Simple covariances and correlation coefficient have proved to be effective to find boreholes apparently connected, but several limitations have been identified:

- **Common forcing**. All boreholes somehow connected to surface water inputs will show a high correlation as long as their pressure variations are in phase. Although, the similarity of pressure variations due to a shared forcing is an issue affecting all the methodologies we tried, the covariances performance is particularly poor at resolving distinct drainage systems.

- **Amplitude data loss**. Normalized covariances (correlation coefficients, equation S2) are useful to detect connections between boreholes with very different oscillation amplitudes, but they discard useful amplitude information as efficient connections should lead to similar amplitudes between boreholes.

- **Time shifts**. Small time shifts as expected for diffusive hydraulic connections or datalogger clocks drift can hide strong correlations. Cross-covariances could be an option to find maximum covariances at different offsets within a reasonable range, but it would be more intensive computationally.

In conclusion, an approach based only in the covariance is not sensitive enough to the different properties that make two time series similar or not. Nevertheless, improvements to overcome those limitations can be envisioned. For example, by defining a "connectivity index" that combines the covariances between different pre-processed versions of the pressure signal, such as the envelope and the diurnal residuals.

S2.3 Self Organizing Maps (SOMs)

Self-Organizing Maps is a form of neural network that uses competitive learning to produce a topologically accurate representation of a dataset reducing its dimensionality, becoming a very powerful unsupervised clustering algorithm. SOM is an iterative algorithm that for given numbers N and M, creates a "map" consisting of $N \times M$ nodes organized in a two-dimensional array. Each node or "unit" contain a time series that is optimized to minimize the Euclidean distance (minimum squares), to a subset of time series in the data set. In this way, the final map, contain the $N \times M$ set of time series that better represent the data, and each data series (boreholes in our case) can be assigned to one map unit. In this way, similar time series would likely be grouped in the same unit. An advantage of this method is that it can easily be applied independently or simultaneously to the raw data, and any set of pre-processed versions of it.

SOM clustering was very successful finding similar data series. However, the size of the map has to be predefined, a requirement that turned out to be very problematic for our application. For example, if a reasonable map size is adopted based on the number of sensors for a given time window, if there are abundant connections (as is common in early spring), the map would be oversized, and it could force the separation of one coherent cluster into multiple groups. On the other hand, in a typical winter situation with many isolated sensors, there will not be enough map units to capture the diversity of the system, and if there is a coherent cluster, it will likely be contaminated with other unrelated sensors.

We envisioned and attempted few approaches to an adaptive map size, finding iteratively the size that minimizes quantization and topological errors, but this, together with all the tunable parameters of the SOM algorithm made the approach cumbersome and computationally expensive. For this reason, before committing to a highly specialized SOM clustering approach, we decided first to explore other simpler techniques that might produce equivalent or better results.

Figure S2 show a map with the 53 sensors operating over ten days, were distinct clusters were captured in units #4, #7 and #10. An example of a group catching unrelated sensors is #6, suggesting that the map is too small. However, if recomputed on a larger size, group #10 would be split into two map units, requiring re-clustering.

S2.4 K-means clustering

K-means clustering is a very popular and perhaps the simplest clustering algorithm. In combination with a standard Euclidian metrics, it was used by Fudge et al. (2008) on a small array of 16 boreholes on Bench Glacier, Alaska. However, as with SOM's it requires the number of clusters to be defined beforehand, running in the same kind of problems described above. For this reason, it was not tested, given that hierarchical clustering was a more promising technique.

S2.5 Hierarchical clustering

Described in the main text.

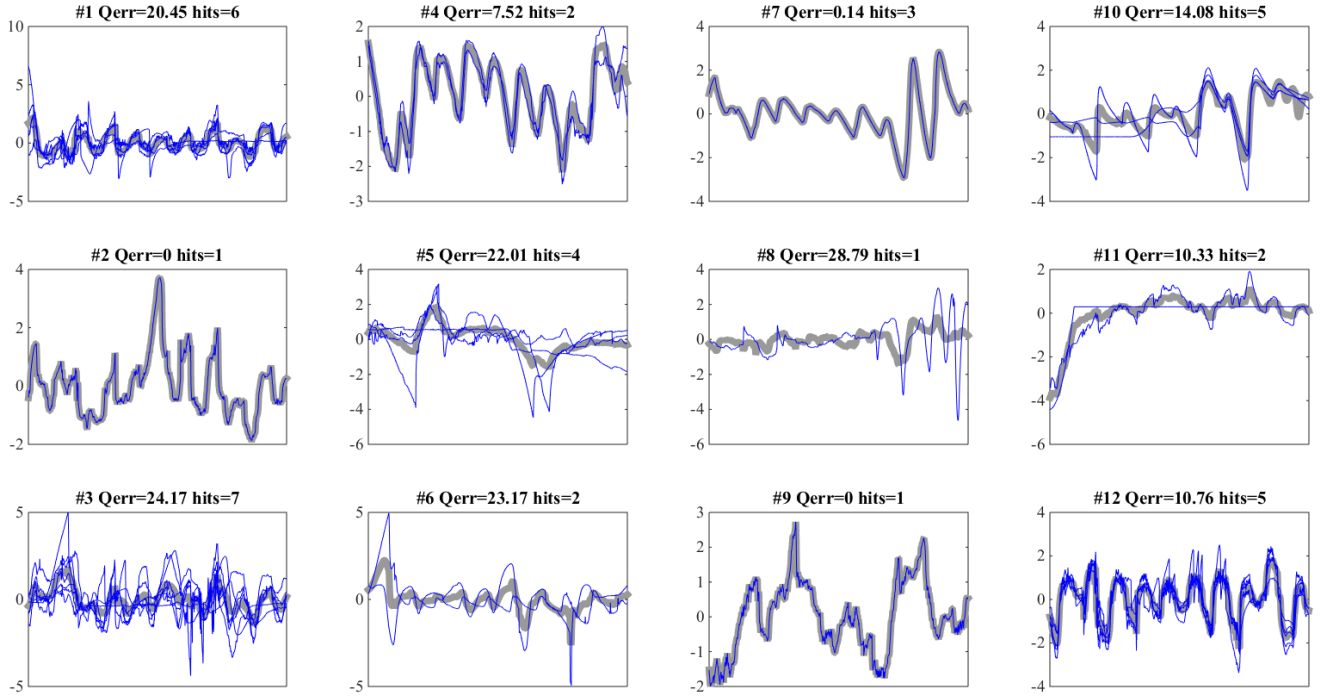

Figure S2. Self-Organizing Map (SOM) of 3 by 4 units in size computed from the pressure residuals of 53 sensors operating between July 28^{th} and August 8^{th} 2011. The time series representative of each unit (thick gray line) is shown together with the sensors matching the unit. For each unit, an average quantization error is shown (average Euclidean distance between each line and the one representative of the unit) and the number of hits (number of sensors matching that unit). The overall quantization error of the map is 15.78, and the topological error is 0.13.

S3 Distance metrics

The clustering technique presented in the main text relies in the the absolute Euclidean distance metrics (See Eq. 2). Here, we present a detailed account and formulation of each of the alternative distance metrics tested during the development of the clustering technique. These distance metrics are

- **Euclidian**: Differences between time series a and b, each one with N samples are measured as Euclidian distance in an N-dimensional space.

$$D_E(a,b) = \sqrt{\sum_i^N (a_i - b_i)^2} \tag{S7}$$

- **Correlation**: Time series differences are quantified by the covariance complement. Therefore, distance is small for highly correlated signals. It is worth mentioning that for standardized time series, the Correlation and Euclidian distances are mathematically equivalent.

$$D_C(a,b) = 1 - \frac{cov(a,b)}{\sigma_a \sigma_b} = 1 - \frac{\sum_i^N (a_i - \bar{a})\left(b_i - \bar{b}\right)}{\sqrt{\sum_i^N (a_i - \bar{a})^2}\sqrt{\sum_i^N \left(b_i - \bar{b}\right)^2}} \tag{S8}$$

where $cov(a,b)$ is the covariance between series a and b, σ_a and σ_b are their respective standard deviations and $\bar{a}$, $\bar{b}$ their mean values.

– **Absolute correlation**: Time series differences are quantified by the absolute covariance complement. This produce small distance between well correlated and also anti-correlated signals.

$$D_{AC}(a,b) = 1 - \frac{|cov(a,b)|}{\sigma_a \sigma_b} = 1 - \frac{\left|\sum_i^N (a_i - \bar{a})\left(b_i - \bar{b}\right)\right|}{\sqrt{\sum_i^N (a_i - \bar{a})^2}\sqrt{\sum_i^N \left(b_i - \bar{b}\right)^2}} \tag{S9}$$

where $|x|$ stands for the absolute value of x.

– **Dynamic Time Wrapping**: DTW is a standard technique used in voice recognition systems, allowing to measure the similarities between waveforms but insensitive to differences and changes in speed between the reference waveform and the one being classified (Mullin, 1983). In our context, DTW can help recognizing similar signals that have a small time shift. Such time shift can arise from datalogger clock offsets or physical processes such as diffusion of pressure signals.

The following pseudo-code describe the used DTW algorithm

a, b Vectors of size N containing the two time series
w Maximum offset allowed as number of samples
D Matrix of size $(N+1) \times (N+1)$ fill up with *infinites*
$D(1,1) = 0$

for $i = 1$ to N
  for $j = max(i-w, 1)$ to $min(i+w, N)$
    distance $= (a(i) - b(j))^2$
    $D(i+1, j+1) =$ distance $+ min\left(D(i, j+1), D(i+1, j), D(i, j)\right)$
  end j loop
end i loop
$D_{DTW} = D(N+1, N+1)$

– **Absolute Dynamic Time Wrapping**: The minimum between the standard DTW distance (as described before), and the DTW distance to the reversed time series (i.e. up/down flipped). This allows detection of anti-correlated sensors.

$$D_{ADTW}(a,b) = min\left(D_{DTW}(a,b), D_{DTW}(a,-b)\right) \tag{S10}$$

S4 Clustering performance evaluation: Entropy and Information Gain

In order to assess the performance of the clustering algorithms and distance metrics described above we have used the information gain achieved by each alternative. Information gain is a standard quantity used in decision tree analysis (Mitchell, 1997): when elements belonging to two classes are grouped according to their class, the information gain measures the quality of the grouping. The Relative Information Gain (RIG) is the information gain relative to the maximum possible gain, which is

achieved when each element is correctly assigned to the group it belongs. We can represent the RIG as a percentage: a RIG of 100% means the information gain was maximum, and the clustering strategy reached its maximum possible performance (i.e. all elements were assigned to the correct group).

If we have a group G made out of N objects that belong to M classes $C_1, C_2, ...C_M$, the entropy E of a group of them, is a measure of how "pure" is the group and is given by

$$E(G) = -\sum_{i=1}^{M} \frac{N_{C_i}}{N} \log\left(\frac{N_{C_i}}{N}\right) \tag{S11}$$

where N_{C_i} is the number of objects belonging to the i^{th} class. If we have only two classes **A** and **B**, and we call f_A and f_B to the fraction of the objects that belong to each class respectively, then the entropy will be

$$E(G) = -f_A \log(f_A) - f_B \log(f_B) \tag{S12}$$

where two relevant extremes are $f_A = f_B$ in which case we have the maximum possible entropy $E = 1$, or when either f_A or f_B is zero, then we have only one class and minimum entropy $E = 0$.

This two-classes case is the one we have if we classify all time series in the pressure dataset for a given time window as either belonging to a cluster or not. Lets call this two classes $\mathbf{C}$ for "connected" or $\mathbf{U}$ for "unconnected". If a given clustering strategy splits the dataset G into two sub-groups: $\mathbf{IN}$ for the ones identified as connected to a single system and $\mathbf{OUT}$ for the one unconnected to that particular system, the entropy of the dataset after the splitting G' will be

$$E(G') = f_{\mathbf{IN}} E(\mathbf{IN}) + f_{\mathbf{OUT}} E(\mathbf{OUT}) \tag{S13}$$

Now the effectiveness of the clustering strategy for separating the dataset into sub-groups purely composed by connected or unconnected sensors can be measured by the entropy loss or **information gain** ($\mathbf{Gain}$) given by

$$\mathbf{Gain} = E(G) - E(G') = E(G) - \Big(f_{\mathbf{IN}} E(\mathbf{IN}) + f_{\mathbf{OUT}} E(\mathbf{OUT})\Big) \tag{S14}$$

Similarly, the relative information gain (RIG) can be defined as

$$\text{RIG} = \frac{E(G) - E(G')}{E(G)} = 1 - \frac{\Big(f_{\mathbf{IN}} E(\mathbf{IN}) + f_{\mathbf{OUT}} E(\mathbf{OUT})\Big)}{E(G)} \tag{S15}$$

which can be represented as a percentage: a RIG of 100% means the information gain was maximum, and the clustering strategy reached its maximum possible performance, putting all sensors connected to a particular system in the $\mathbf{IN}$ sub-group.

S5 Hierarchical clustering variations assessment

The performance of agglomerative hierarchical clustering to group hydraulically connected boreholes was evaluated using each of the pre-processing procedures described in Section S2 and the distance metrics described in Section S3. We assessed the performance using the achieved Relative Information Gain (RIG) described in Section S4. The performance was evaluated independently for both hydraulic and Machanical clusters, as well as over all clusters together. This later approach proved to be the most effective.

S5.1 Hydraulic clusters

The best performing method reached an average Relative Information Gain (RIG) of 84%, and used absolute euclidean (correlation) distance metric on the diurnal residual of pressure time series and average-link linkage with a depth based Optimal Split Point (OSP) of 27.174.

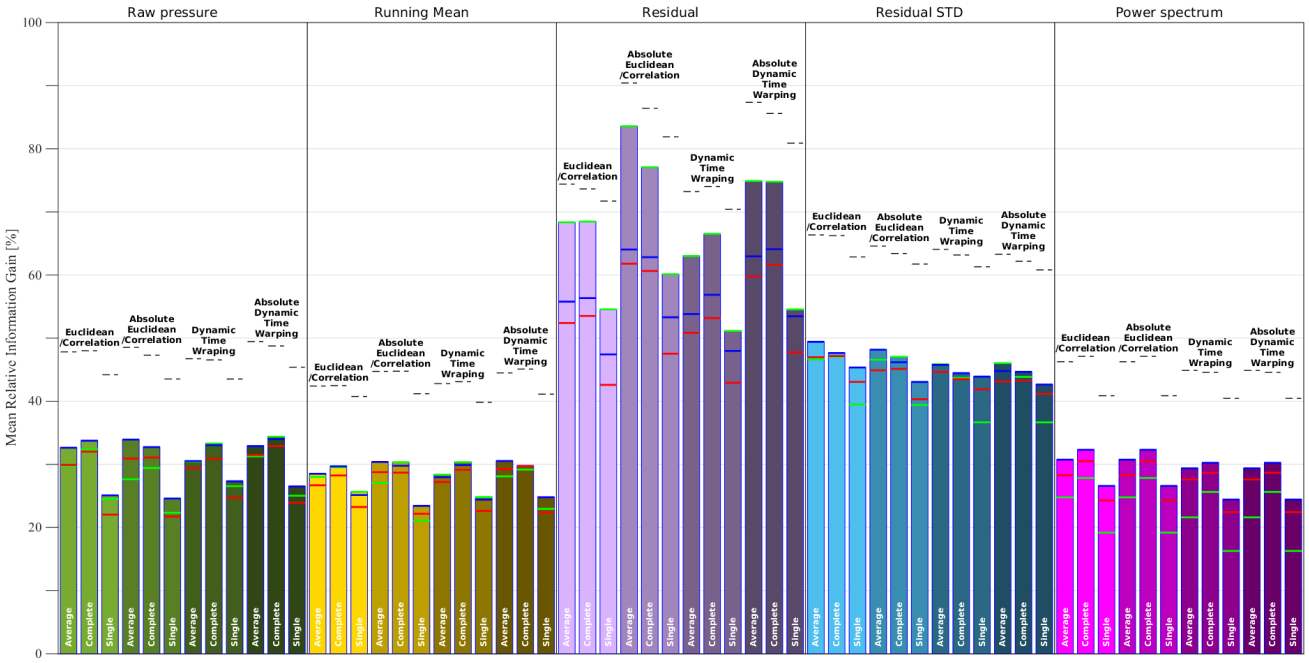

Figure S3. Mean Relative information Gain (RIG) achieved over the 156 hydraulic clusters of the calibration dataset by each tested methodology. Dashed black lines show the maximum RIG that a perfect dendrogram split algorithm could reach on each case. The actual mean RIG achieved by each split method is shown for the best performing Depth OSP (green), Inconsistency OSP (Blue) and Gap OSP (red).

Using the OSPs computed from the calibration dataset, the performance was evaluated on the validation dataset. Figure S4 show the results, on which again the best performing method was absolute Euclidean (correlation) distance metric on diurnal residual of pressure and average-link linkage with a depth based Optimal Split Point (OSP). The mean RIG was 85%, a value consistent with the performance observed during the calibration.

Performance was finally assess on the test dataset, showing again that the best performance was achieved by the method using absolute euclidean (correlation) similarity metric on diurnal residual of pressure and average-link linkage with a depth based Optimal Split Point (OSP). Using the same OSP computed during the calibration, the mean RIG was 87%, consistent with the performance observed during the calibration and validation.

S5.2 Mechanical clusters

The best performing method reached and average Relative Information Gain (RIG) of 89%, and used absolute euclidean (correlation) distance metric on diurnal residual of pressure time series and average-link linkage with a depth based Optimal Split Point (OSP) of 27.097.

Using the OSPs computed from the calibration dataset, the performance was evaluated on the validation dataset. Figure S7 show the results. The best performing method during calibration (absolute Euclidean distance metric on diurnal residuals of pressure and average-link linkage with a depth based OSP) showed the best performance again with a mean RIG of 86%.

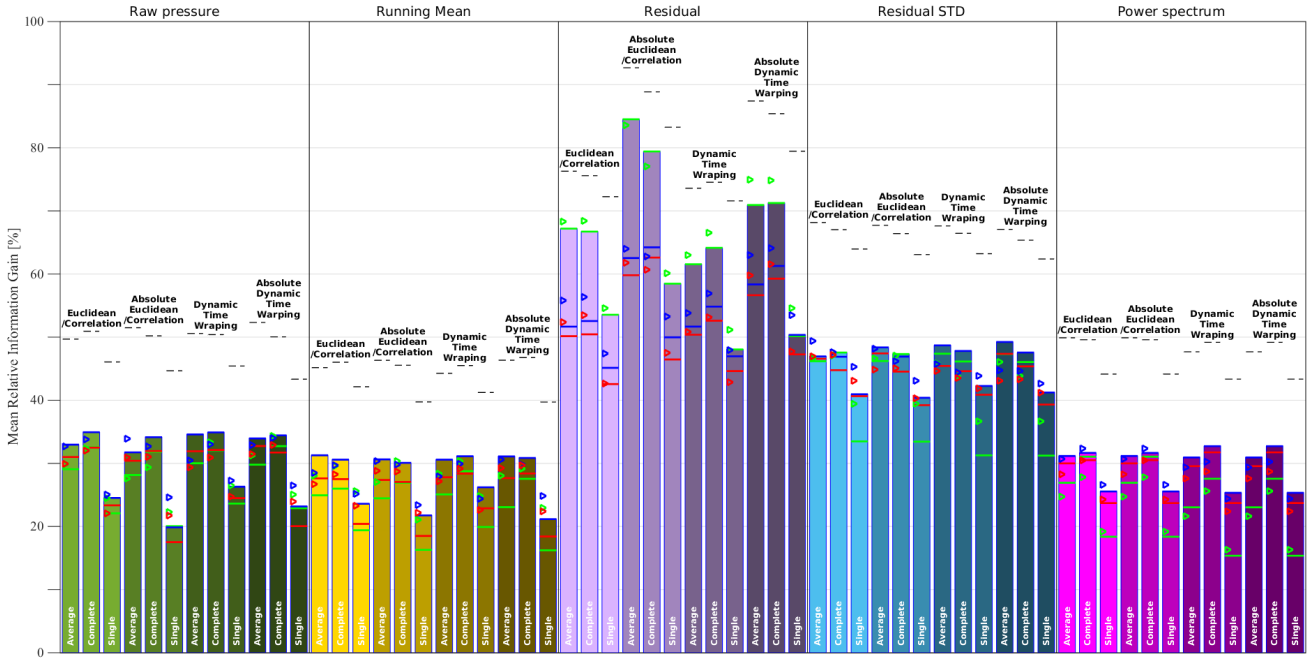

Figure S4. Mean Relative information Gain (RIG) achieved over the 148 hydraulic clusters of the validation dataset by each of the best performing methodologies found during the calibration. Colored triangles point to the reference RIG obtained on the calibration dataset. Triangles and OSP lines follow the same schema as in Fig. S3.

Performance was finally assess on the test dataset, showing again that the best performance was achieved by the method using absolute Euclidean distance metric on diurnal residuals of pressure and average-link linkage with a depth based OSP. The mean RIG was 78%, outperforming again other methods, consistent with the performance observed during the calibration.

S5.3 All clusters

If all clusters are considered without distinction. The calibration dataset results also reports absolute Euclidean distance metric on diurnal residual of pressure and average-link linkage with a depth OSP (of 27.174) as the best clustering method, as shown in Fig. S9. Note that the OSP found is the same than for the hydraulic-only dataset. Therefore, unifying the clustering for mechanical and hydraulical clusters would only reduce the performance on mechanical ones.

Using this OSP to test performance only on mechanical clusters on the calibration dataset leads tho a RIG of 89%, therefore the performance is the same than for the mechanical optimized OSP (27.097). In the validation dataset the performance stay high at a RIG of 86%, and in the test dataset it is 78%. These results suggest that applying a different clustering technique to hydraulical and mechanical clusters does not produce any advantage. Therefore, we will apply only one clustering technique.

Using the OSP found during the calibration with all clusters (Fig. S9), the performance was evaluated on the validation dataset. Figure S10 show the results, on which the best performing method during calibration (absolute Euclidean distance metric on diurnal residuals of pressure and average-link linkage with a depth based OSP) reached a mean RIG of 85%.

Performance was finally assess on the test dataset (Fig. S11), showing again that the best performance was achieved by the method using absolute Euclidean distance metric on diurnal residuals of pressure and average-link linkage with a depth OSP. The mean RIG was 84%.

Summarizing, the best performing hierarchical clustering technique is to use absolute Euclidean distance as similarity metric (equivalent to correlation) on diurnal residuals of pressure time series and average-link linkage with a depth based Optimal Split Point (OSP) of 27.174. A method that achieved a 85%, 85% and 84% on the calibration, validation and test datasets.

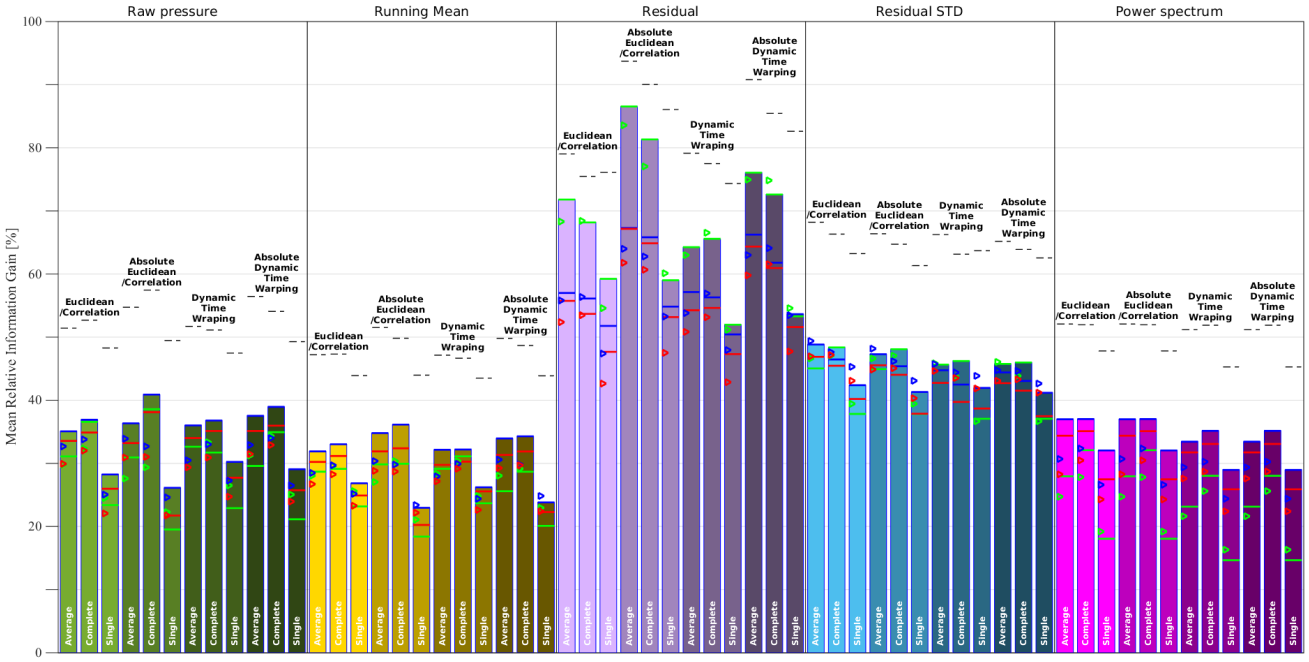

Figure S5. Mean Relative information Gain (RIG) achieved over the 152 hydraulic clusters of the test dataset by the best performing methodologies found during the calibration and validation. Reference RIG triangles and OSP lines follow the same schema as in Fig. S4.

References

David, A. and Vassilvitskii, S.: K-means++: The Advantages of Careful Seeding, in: SODA 07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, 2007.

Fudge, T., Humphrey, N., Harper, J., and Pfeffer, W.: Diurnal fluctuations in borehole water levels: configuration of the drainage system beneath Bench Glacier, Alaska, USA, journal of Glaciology, 54, 297–306, 2008.

Jolliffe, I.: Principal Component Analysis, Springer, 2nd ed. edn., 2002.

Mitchell, T.: Machine Learning, McGraw Hill, Oxford, New York, 1997.

Mullin, R.: Time warps, string edits, and macromolecules: The theory and practice of sequence comparison, Canadian Journal of Statistics, 13, 167–168, https://doi.org/10.2307/3314879, 1983.

Rokach, L. and Maimon, O.: Clustering methods. Data mining and knowledge discovery handbook, Springer, 2005.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J.: SOM Toolbox for Matlab 5, Tech. rep., Helsinki University of Technology, 2000.

Zahn, C. T.: Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters, IEEE Transactions on Computers, C-20, 68–86, https://doi.org/10.1109/T-C.1971.223083, 1971.

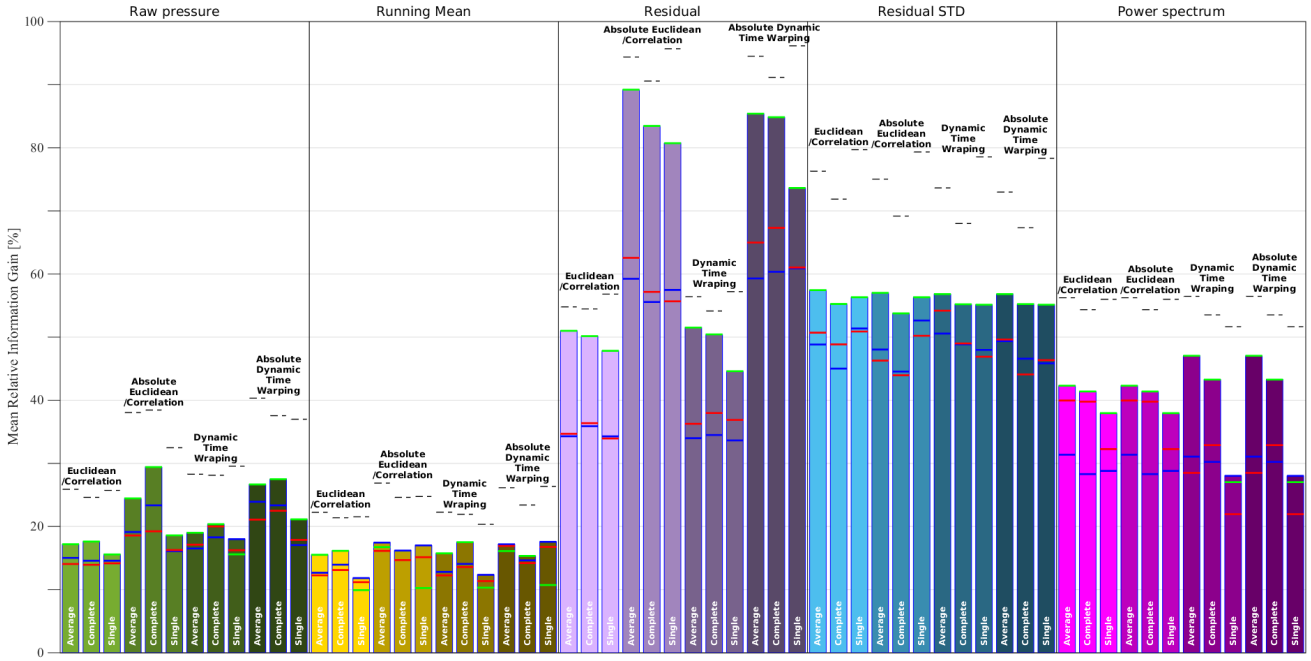

Figure S6. Mean Relative information Gain (RIG) achieved over the 49 mechanical clusters of the calibration dataset by each tested methodology. OSP lines follow the same schema as in Fig. S3.

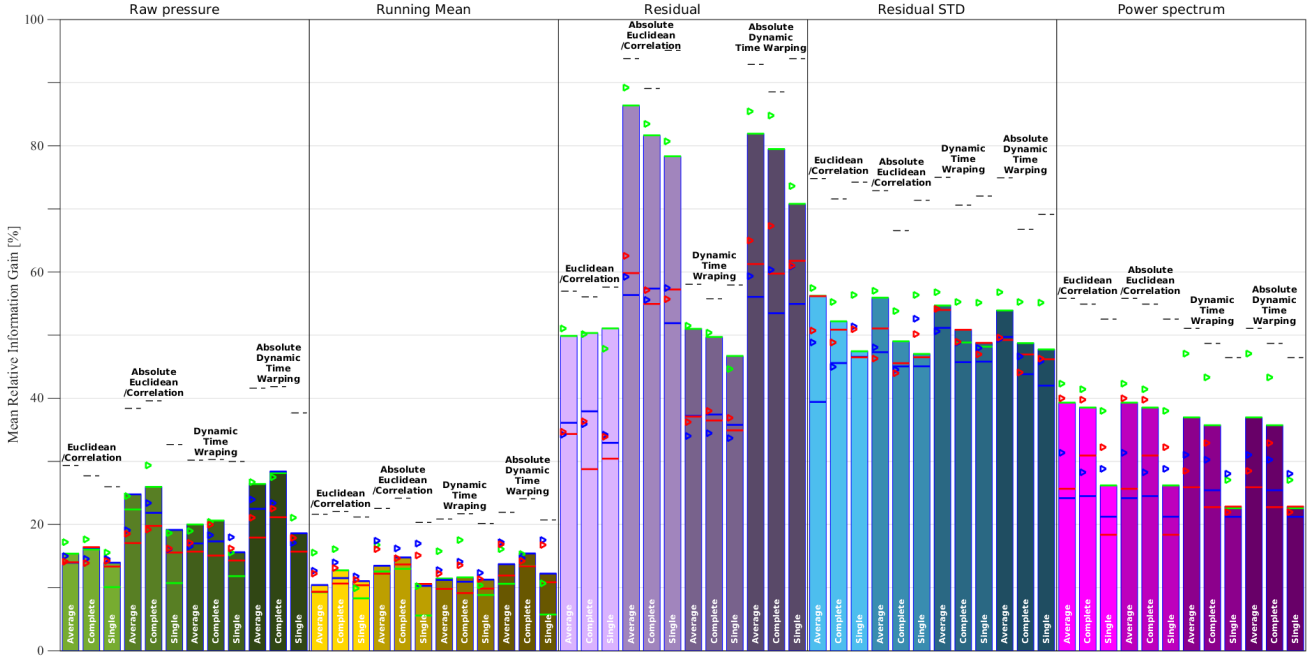

Figure S7. Mean Relative information Gain (RIG) achieved over the 55 mechanical clusters of the validation dataset by each of the tested methodologies. Reference triangles and OSP lines follow the same schema as in Fig. S4.

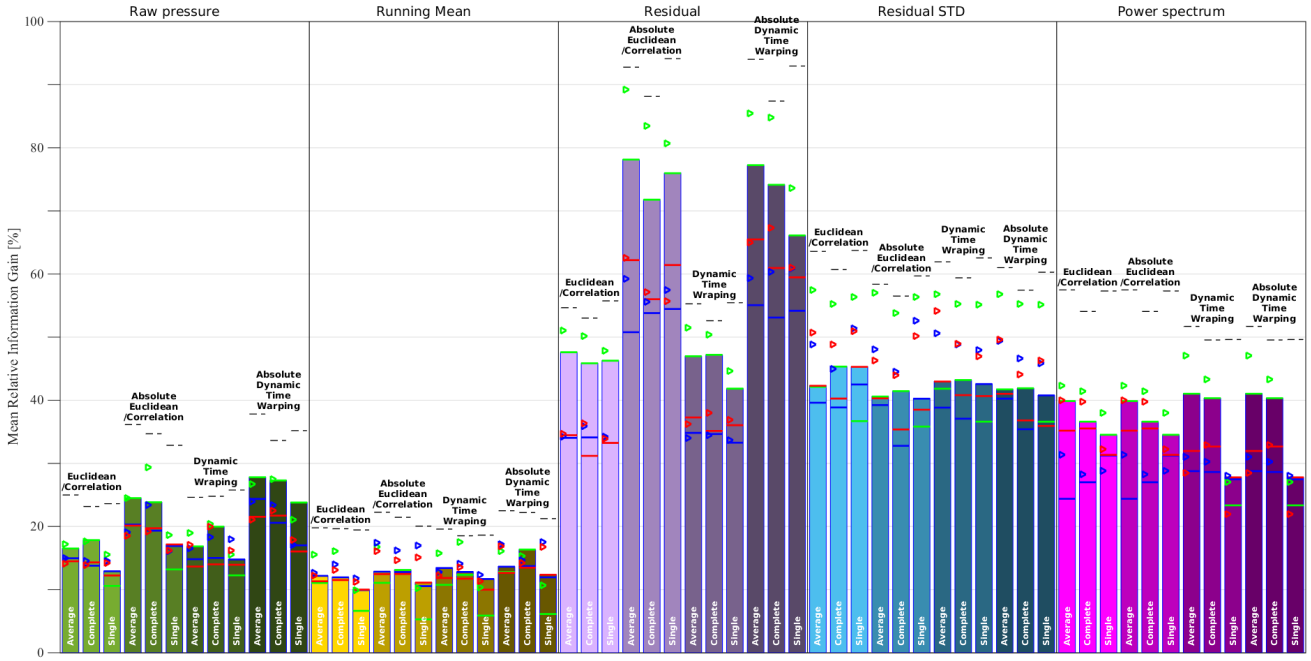

Figure S8. Mean Relative information Gain (RIG) achieved over the 53 mechanical clusters of the test dataset. Reference triangles and OSP lines follow the same schema as in Fig. S4.

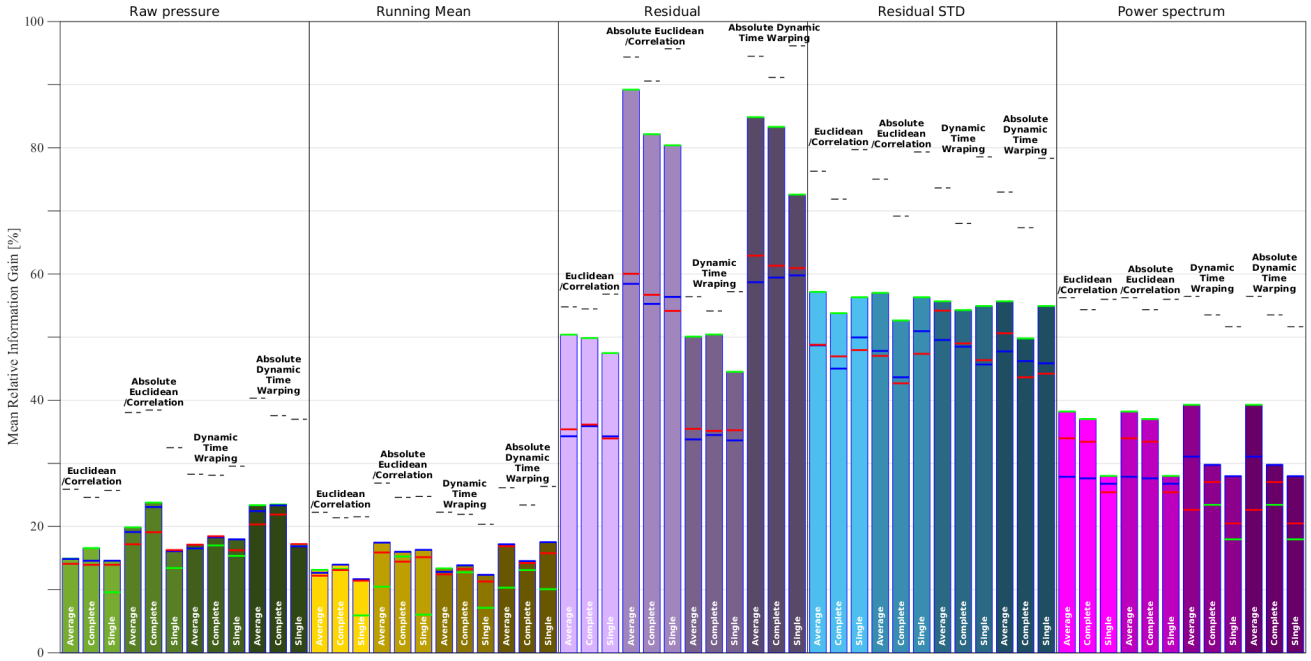

Figure S9. Mean Relative information Gain (RIG) achieved over the 205 clusters of the calibration dataset by each tested methodology. OSP lines follow the same schema as in Fig. S3.

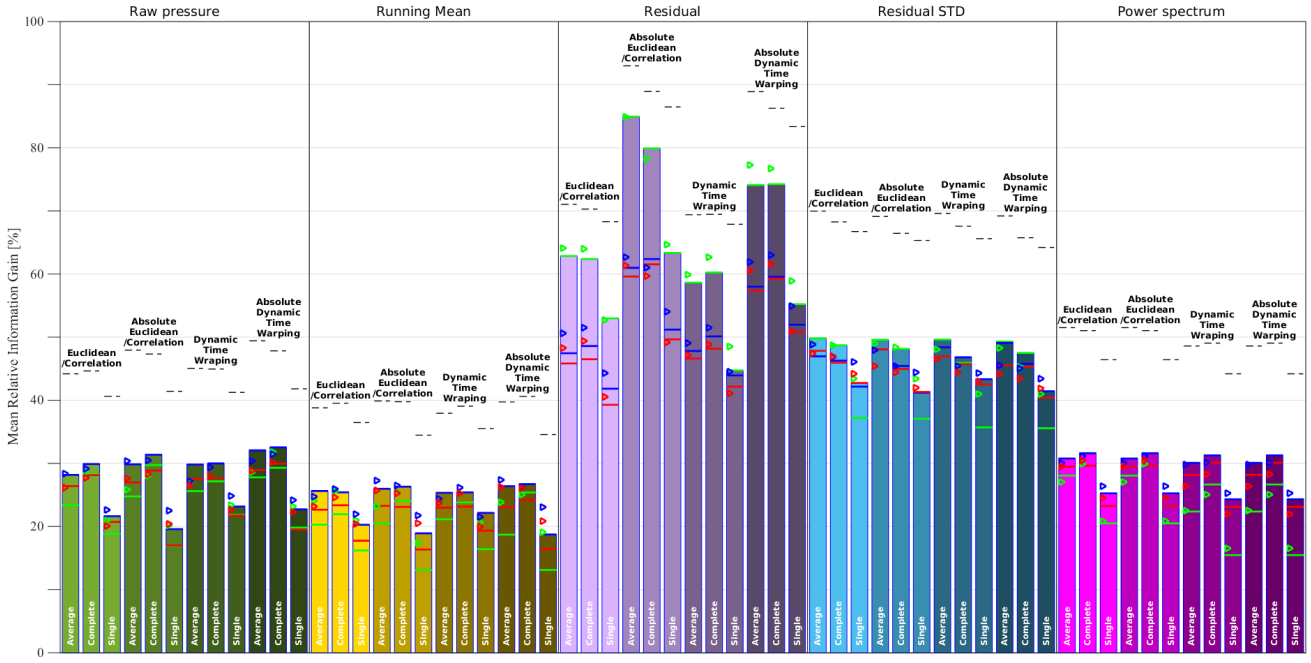

Figure S10. Mean Relative information Gain (RIG) achieved over the 203 clusters of the validation dataset. Reference triangles and OSP lines follow the same schema as in Fig. S4.

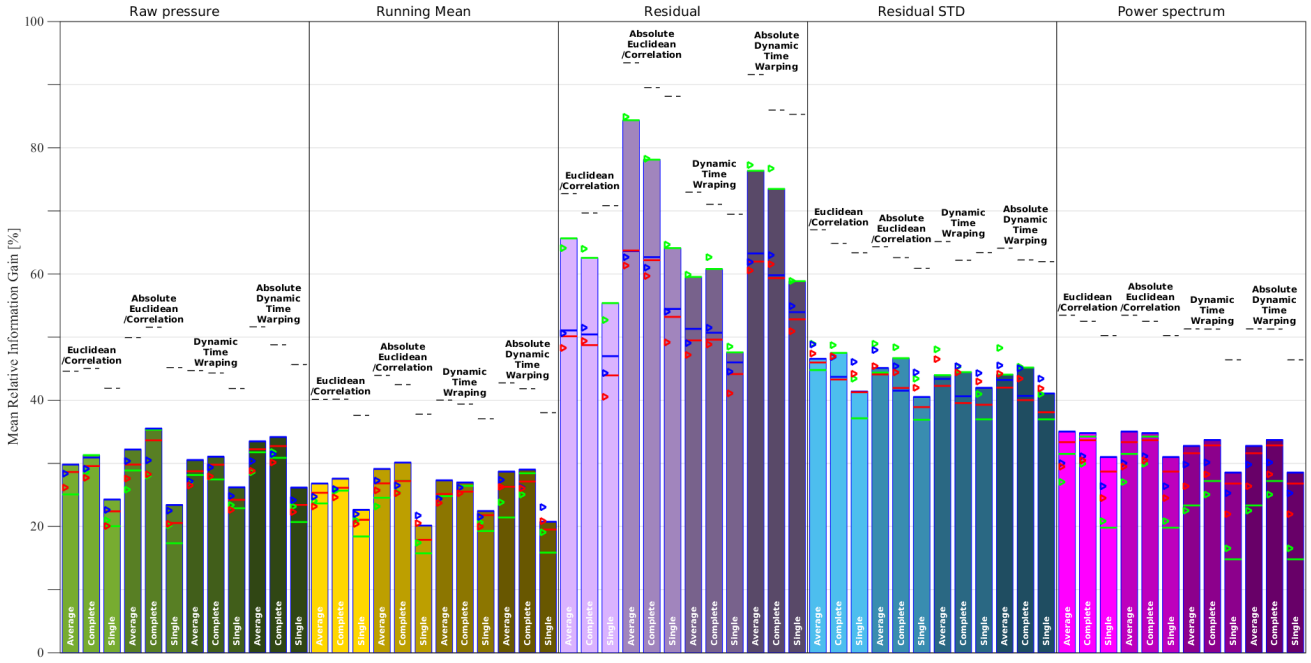

Figure S11. Mean Relative information Gain (RIG) achieved over 205 clusters of the test dataset. Reference triangles and OSP lines follow the same schema as in Fig. S4.