Technical Report: Measuring fidelity of DTs –Incubator–

Paula Muñoz, Javier Troya, Antonio Vallecillo
ITIS Software
University of Málaga (Spain)

Manuel Wimmer
CDL-Mint
Johannes Kepler University (Austria)

Contents

I Incubator in the Aarhus University

A. System description

At Aarhus University in Denmark, they are building a digital twin for their studies in the field, using an incubator as the foundation case study. They built this system, shown in Figure 1, as a simple yet representative thermal incubator system. The main goal of the digital twin system of the incubator is to reach a certain temperature within a box and regulate it regardless of the content inside.

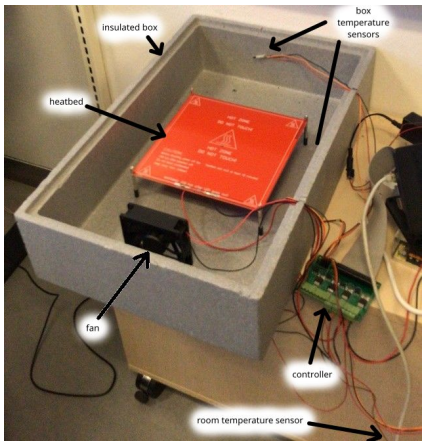


Fig. 1: Incubator in the University of Aarhus (Denmark)

The physical twin of the incubator includes a set of components which form a plant that is controlled by a Raspberry Pi. The system includes an insulated box fitted with a heatbed, and complete with a software system for communication, a controller, and simulation models. The system utilizes two DHT22 sensors (available at https://components101.com/sensors/dht22-pinout-specs-datasheet) to monitor the temperature inside the incubator, while another sensor measures the room temperature. The sensors are programmed to provide readings once every two seconds. The controller, which operates every three seconds, calculates the average temperature readings from the two sensors inside the incubator to determine its internal temperature.

Two simulation models were developed to act as the digital twin. They can predict the temperature inside the incubator, one with two free parameters and one with four free parameters.

1) 2-parameter model

The two parameter model for the incubator is a simple linear model which predicts the temperature inside using two parameters: the heat transfer coefficient and thermal resistance. It assumes that these parameters remain constant throughout the operation. To determine the values of these parameters, experimental data from the physical incubator system was used for calibration. The calibration process involved minimizing the difference between the predicted temperatures from the model and the actual measured temperatures using a least-squares method. The two-parameter model was found to be accurate in predicting the incubator's temperature. However, it should be noted that the model has limitations in capturing non-linear dynamics and transient behavior of the system.

2) 4-parameter model

The four-parameter temperature prediction model for the incubator is an advanced non-linear model designed to capture the system's non-linear dynamics and transient behavior. This model considers the heat transfer coefficient, thermal resistance, thermal capacitance, and a time constant as its key parameters. To determine the optimal values for these parameters, experimental data from the physical incubator system was used for calibration. The calibration process involved minimizing the discrepancy between the predicted temperatures from the model and the actual measured temperatures using a least-squares method, as in the two-parameter model. The four-parameter model exhibited greater accuracy compared to the simpler two-parameter model, particularly during transient behavior. However, it comes with the trade-off of requiring more computational resources and being more complex in nature.

More detailed information about the incubator system description and implementation are available in [1] and the complete implementation of the system in [2].

B. Scenarios

The objective of the incubator is to maintain the temperature within specific boundaries. For the scenarios we are working with, the upper limit will be set at 35 degrees Celsius, which is the desired temperature, while the lower limit will be 30 degrees Celsius. As a result, the temperature inside the incubator will fluctuate within these limits.

1) Heating time 3 s - Heating gap 2 s (Ht3Hg2)

In Figure 2, we can observe the heating and cooling patterns for the first scenario: *Heating time 3 s - Heating gap 2 s (Ht3Hg2)*. The heating pattern is within the temperature range of [35, 30] degrees Celsius for the physical twin and the two digital twins of the incubator. In this particular scenario, the heater is activated during the heating process, following a pattern of heating for 3 seconds, turning off for 2 seconds, and then reactivating if the controller detects that the target temperature has not been reached. These short decision-making periods help maintain the temperature within tight boundaries, ensuring that the system consistently operates within the specified range.

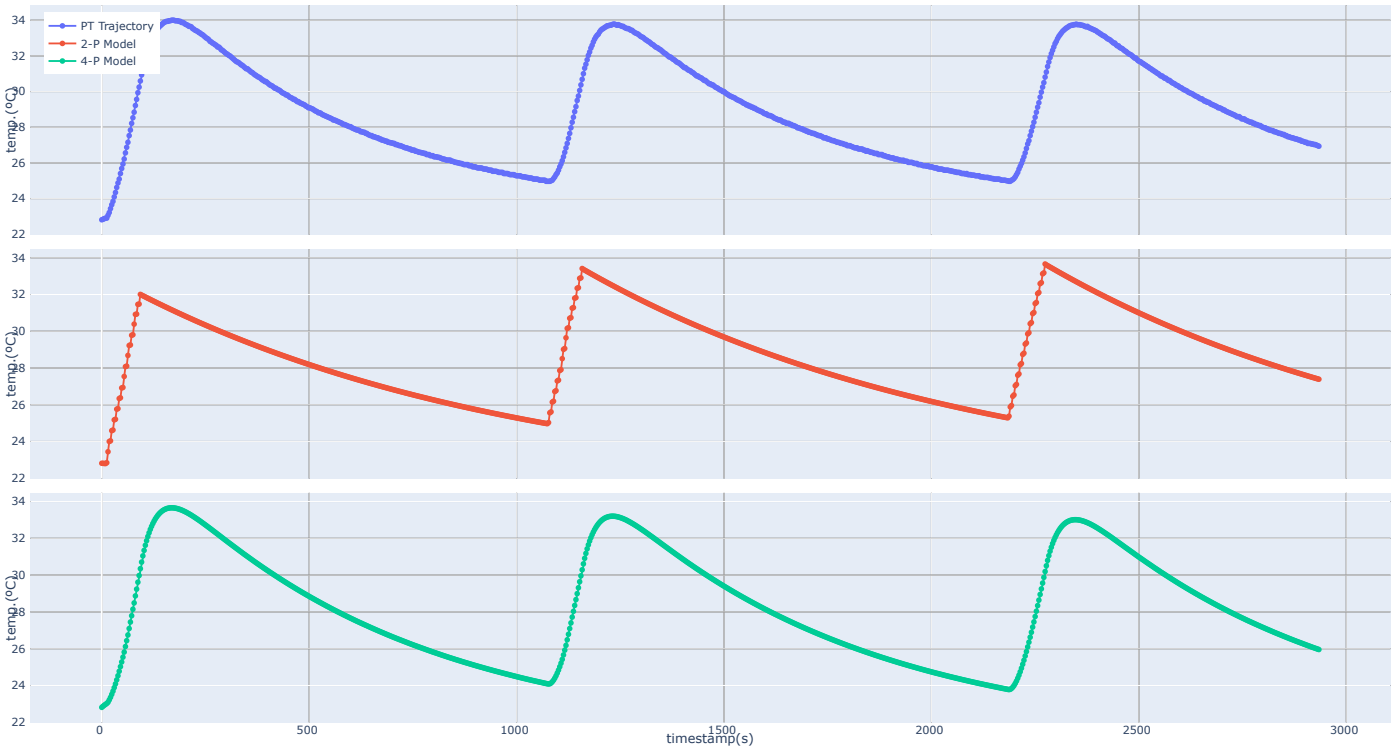


Fig. 2: Traces for scenario Ht3Hg2 (In order: Physical Twin, 2-P model, 4-P model)

2) Heating time 30 s - Heating gap 20 s (Ht30Hg20)

In Figure 3, we can observe the heating patterns of the second scenario: *Heating time 30 s - Heating gap 20 s (Ht30Hg20)*. In this scenario, the controller activates the heater for 30 seconds if it detects that the temperature is below the upper limit, and then turns it off for 20 seconds. It reactivates the heater if, after these 20 seconds, the temperature is still below the limit. This spaced-out control intervals result in slower decision-making and cause the system to move outside the specified temperature limits, leading to poorer performance based on the problem parameters. Additionally, this pattern creates a sawtooth pattern in the two-parameter model, which is unable to smooth out the temperature changes caused by this control scheme.

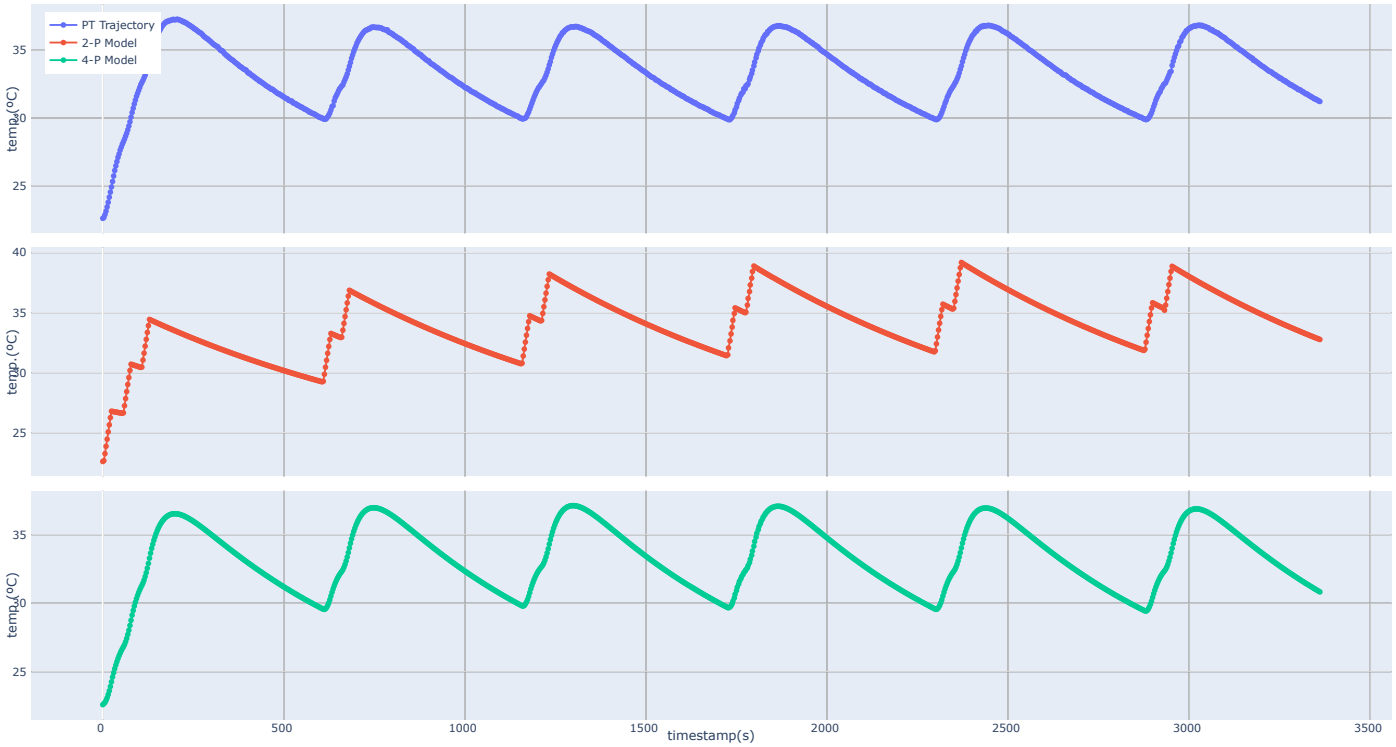


Fig. 3: Traces for scenario Ht30Hg20 (In order: Physical Twin, 2-P model, 4-P model)

C. Gap Tunning

Verifiability: The analysis performed in this section is available at https://github.com/atenearesearchgroup/fidelity-measure-for-dts/blob/main/src/evaluation/incubator/incubator_gap_tuning.ipynb

One of the main configuration parameters for alignment algorithms is the scoring system. In order to discern among the many possible alignments between two sequences, it is necessary to specify to the algorithm which decisions to prioritize when aligning the sequences. The classical Needleman-Wunsch algorithm, on which our proposal is based, has two configurable penalties (mismatch and gap) and one reward (match). These values can be assigned based on an input matrix that prioritizes certain characters over others (in the original approach for protein sequence alignment) or through fixed values.

One of the most common configurations when using fixed values is to assign +1 for a match, 0 for a mismatch, and -1 for a gap. This type of configuration prioritizes mismatches over gaps, favoring solutions with fewer gaps. In our case, we adapted this scoring configuration:

- **Match:** Value in the range (0, 1]. The more similar the snapshots are, the closer the value is to 1, based on the comparison function.
- **Mismatch:** Neutral penalty, 0. This occurs when the two snapshots fall outside the range of the Maximum Acceptable Difference (MAD). However, the algorithm considers that these snapshots should have matched for the optimal alignment.
- **Gap:** Negative penalty, aimed at prioritizing mismatches over gaps. This represents a state that is absent in the other trace.

This negative penalty can be configured in two ways:

- **Simple gap:** A fixed penalty that is added to the score each time a gap is included in the alignment.
- **Affine gap:** A configuration with two fixed penalties: one penalty for initiating a gap (P_{op}) and another usually smaller penalty for extending a previously initiated gap (P_{ex}).

The first approach produces alignments in which single-position gaps and matches alternate in the sequences. However, this alignment scheme may be less effective when the objective is to identify periods of anomalous behavior, as it tends to result in alignments with intermittent gaps in the trace. Conversely, in the second approach, we introduce penalties for such alignments and instead prioritize alignments where gaps are grouped together. Longer gaps facilitate the identification of anomalies, resulting in more meaningful alignments.

However, the latter approach demands more processing space and computational capacity. It not only requires one matrix to align the sequences using Dynamic Programming but also necessitates two additional matrices to evaluate whether to insert a gap or not in each of the sequences. Hence, in our algorithm, we incorporated the flexibility to configure alignments using both of these techniques. Depending on the specific scenario and the importance given to resource optimization, the user can select either approach. To assess the optimal configurations for penalties and their impact on alignment, we prepared experimental datasets for which we analyze the fidelity metrics introduced in Section III of the General Concepts Technical Report [3].

The configurations for the experiments conducted with the incubator are as follows:

Parameter	**Range**	**Increments**
Maximum Acceptable Distance (MAD)	[0.10, 0.2]	0.02
Penalty opening a gap (P_{op})	[-3.0, 0.0]	0.50
Penalty extending a gap (P_{ex})	[-2.0, 0.0]	0.10

TABLE I: Analysis of the influence of simple gap penalty on the percentage of aligned snapshots.

Model	R-squared	F-statistic	Coef. MAD	P-value MAD	Coef. P_{op}	P-value P_{op}	Coef. P_{ex}	P-value P_{ex}
Simple 2-P (segment)	0.669	144.0	37.844 ± 1.915	0.000	0.728 ± 0.072	0.000	2.856 ± 0.326	0.000
Simple 2-P (all)	0.871	1876.0	87.953 ± 5.959	0.000	4.031 ± 0.204	0.000	25.003 ± 0.353	0.000
Simple 4-P (segment)	0.637	23.940	32.047 ± 4.939	0.000	1.351 ± 0.176	0.000	7.959 ± 1.564	0.000
Simple 4-P (all)	0.962	6985.0	82.027 ± 2.084	0.000	3.660 ± 0.071	0.000	14.985 ± 0.123	0.000

This resulted in an analysis of 840 alignments applied to Scenario Ht3Hg2 for each model (2-P and 4-P). The chosen range of MAD values depends on the system's domain. In the following two sections, we will examine the effects of these configurations on various metrics using figures such as Figure 4a and 4b. In these figures, all subfigures share the x-axis, where each unit represents an alignment applied to the scenario. Depending on the input values (MAD, P_{op}, P_{ex}), we obtain alignments with different statistics. The shading in this and subsequent figures represent change points, dividing the values into groups based on the values' tendency.

1) Simple gap

The statistics for the percentage of matched snapshots, Frèchet distance, and average Euclidean distance between aligned points are depicted in Figures 4a and 4b, for the 2-Parameter model and Figures 5a and 5b, for the 4-Parameter model. These figures specifically focus on the 120 input configurations where the P_{op} value is set to 0, implying that the cost of opening and extending a gap is the same.

The first two figures, 4a and 5a, present the samples sorted along the x-axis based on the increasing percentage of matched snapshots. In the other two figures, 4b and 5b, instead of arranging the samples based on the percentage of matched snapshots, they are sorted in increasing order of the number of gaps in the alignment.

Let us analyze the configuration results for both models.

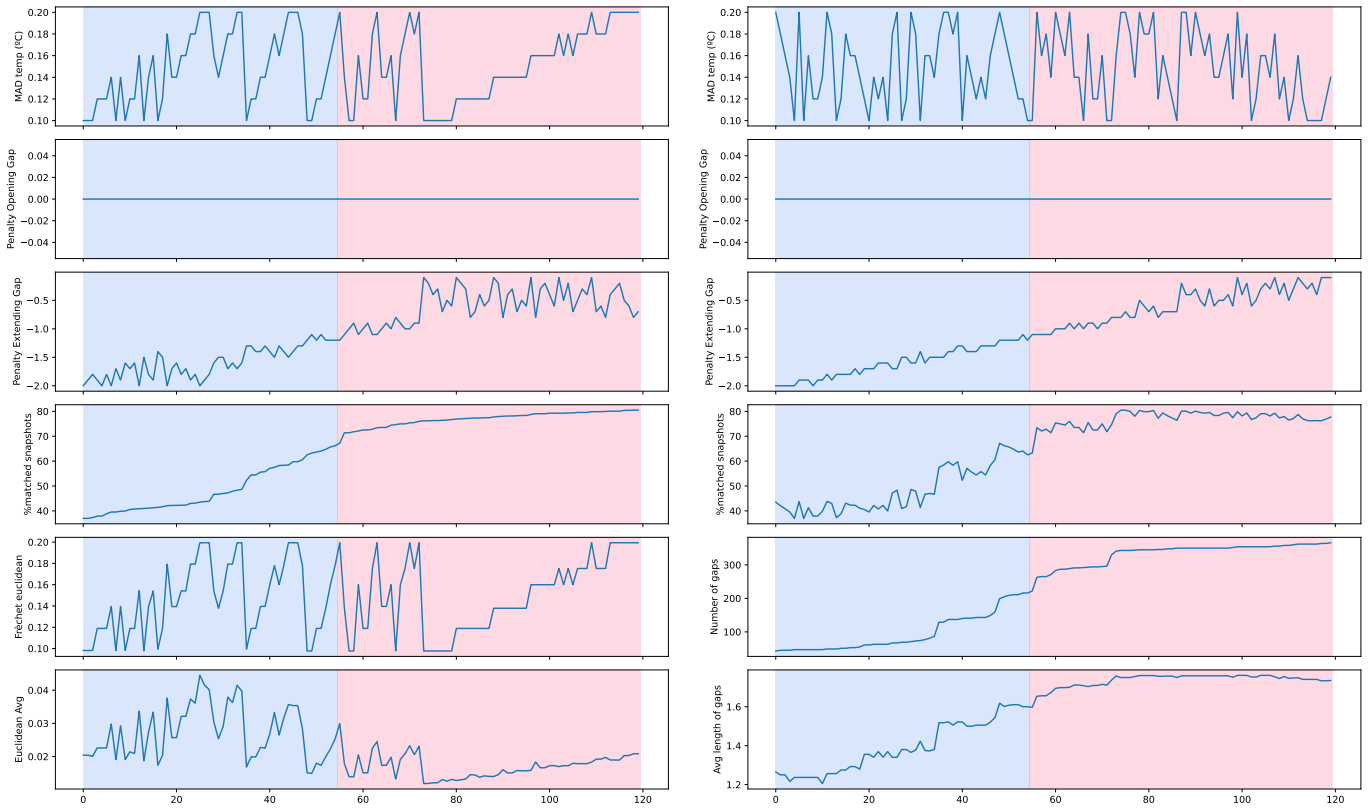
I. 2-Parameters model

The change point, which signifies a shift in the values, is observed at the peak of the percentage of aligned snapshots. This peak includes the samples with the highest percentage of aligned snapshots, which are in the range [70,80] %, shaded in pink in Figure 4a. Within this range, it achieved the highest snapshot percentages and the lowest distances (Frèchet and Euclidean). The Fréchet distance is greatly influenced by the MAD, but the Euclidean Average is the lowest within the aforementioned range, unaffected by the MAD fluctuations. This leads us to believe that these are the best alignments, at least among the ones in which the snapshots are the closest. If we consider the value of P_{ex} in this range, it becomes evident that satisfactory results are only obtained with values greater than -1.0. Values below this threshold imply that the algorithm struggles to provide enough flexibility to incorporate an adequate number of gaps to align snapshots, leading to unsatisfactory outcomes.

We can understand this more easily by looking at Figure 4b. In this case, we can observe a similar change-point as in Figure 4a, obtaining the alignments with the highest percentage of matched snapshots in the range [70, 80]%. Once a certain number of gaps is reached (around 300), the alignments become satisfactory. This happens again for values of P_{ex} greater than -1.0, which means that the penalty for introducing a gap is low enough for the algorithm to prioritize an alignment that includes an adequate number of gaps. These gaps help to characterize the behavior of our system in comparison to the simulation, enabling the inclusion of delays, for example.

To verify the statistical relevance of the input values in relation to the output values, we performed linear regressions that relate the input parameters (MAD, P_{op}, P_{ex}) to the percentage of aligned snapshots. The results of this analysis are available in Table I. In this table, we have the values for the analysis of all samples (Simple 2-P (all)) and specifically for the red-colored segment with the most optimal values (Simple 2-P (segment)). The results for the three input parameters are the following:

- **MAD** has a significant influence on explaining the variability of the data across all samples, as indicated by the statistical relevance with p-values below 0.05 and the coefficients are relatively high: if we change one



(a) Ordered by % of matched snapshots.

(b) Ordered by the number of gaps.

Fig. 4: Analysis of alignment statistics for simple gap for the 2-P Model.

degree, which is a relevant difference in our context, the MAD value, the % of aligned snapshot should change approximately the value of the coefficient which is estimated in 87%. However, by examining Figure 4a, we can observe that the values are distributed throughout the entire range in an increasing fashion. This implies that increasing the MAD loosens the alignment restrictions, resulting in a higher percentage of matched snapshots.

- $\mathbf{P_{op}}$ coefficients are really low for all samples, affecting around 4% for each unit we increase the penalty, which means that it barely affects the percentage of aligned snapshots.
- The $\mathbf{P_{ex}}$ coefficients are high considering all samples, as a change of 1 unit in P_{ex} would increase the percentage of aligned snapshots by 25%. However, when we only consider the aforementioned segment in which we obtain the optimal results, this coefficient is reduced to 2%. This indicates that modifying the value of P_{ex} within the appropriate range of values [-1, 0) has little impact on the percentage of aligned snapshots.

The conclusion is that we can achieve satisfactory results for the incubator example with $\mathbf{P_{ex}}$ **values between [-1.0, 0)**, regardless of the specific variations within that range.

II. 4-Parameters model

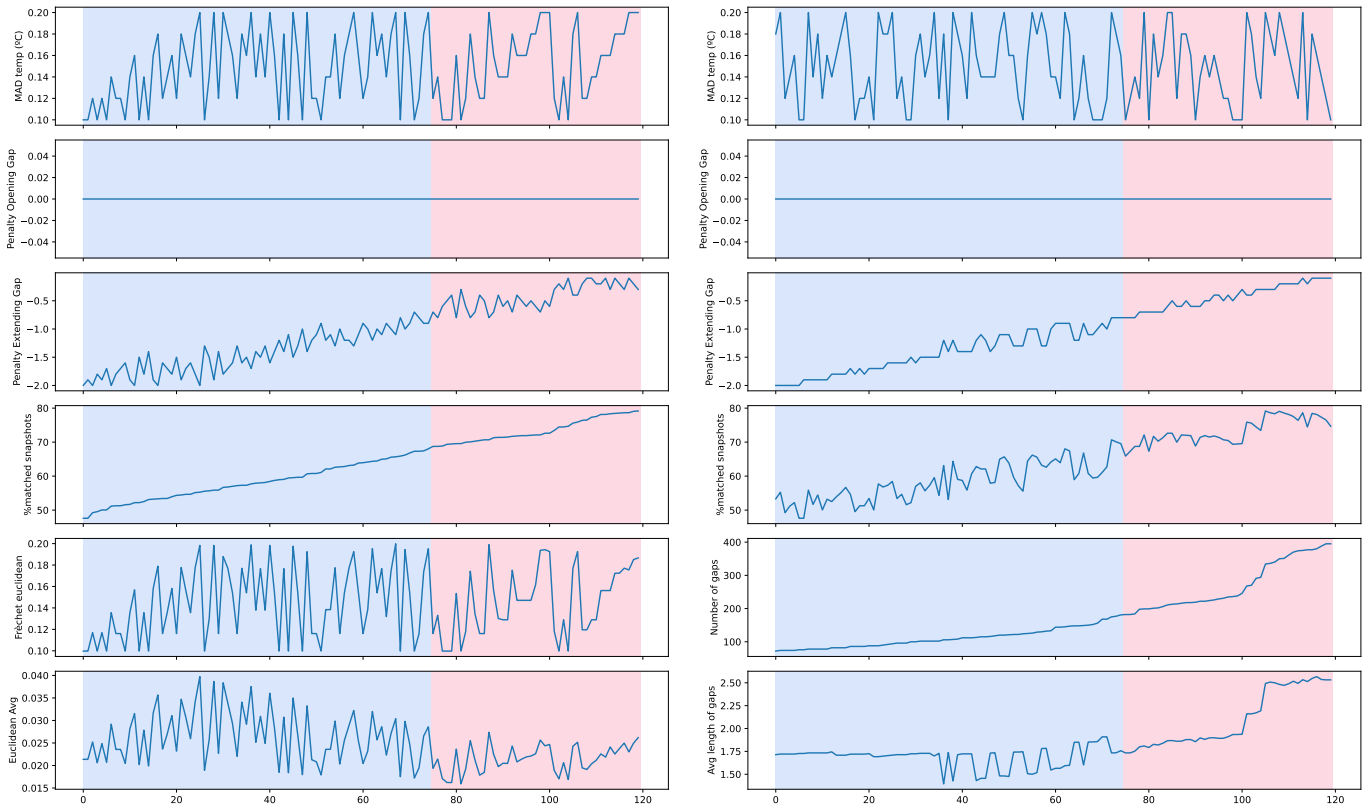
The analysis of the 4-Parameter Model closely resembles that of the 2-Parameter Model. In this case, as shown in Figure 5a, the change point is not as prominent, but there is a shift in trends around 70% of aligned snapshots. The percentage of aligned points progressively increases as we decrease P_{ex}. Similar to the previous case, the best results are achieved with a P_{ex} greater than -1.0. With these configurations, we attain a percentage of aligned snapshots ranging from 70% to 80%. The Fréchet distance is influenced by the MAD, but the Euclidean Average is the lowest within this aforementioned range, unaffected by the MAD fluctuations.

We can further analyze this example more easily by looking at Figure 5b. In this case, we can observe a similar change point as in Figure 5a, obtaining the alignments with the highest percentage of matched snapshots in the range [70, 80]%. Once a certain number of gaps is reached (around 200), with P_{ex} greater than -1.0, the alignments become satisfactory. By further increasing this value up to -0.5, we can achieve an even higher number of aligned snapshots with over 300 gaps in the alignment. This pattern repeats for values of P_{ex} greater than -1.0, indicating that the penalty for introducing a gap is low enough for the algorithm to prioritize an alignment that includes a sufficient number of gaps. These gaps play a crucial role in characterizing the behavior of our system compared to the simulation, allowing for the inclusion of delays, among other factors.

To verify the statistical relevance of the input values in relation to the output values, we performed linear regressions that relate the input parameters (MAD, P_{op}, P_{ex}) to the percentage of aligned snapshots. The results of this analysis are available in Table I. In this table, we have the values for the analysis of all samples (Simple 4-P (all)) and specifically for the red-colored segment with the most optimal values (Simple 4-P (segment)). The results for the three input parameters are the following:

- **MAD** has a significant influence on explaining the variability of the data across all samples, as indicated by the coefficients, which are relatively high: if we change one degree the value of MAD, the % of aligned snapshot should change approximately the value of the coefficient which is between 82%. However, by examining Figure 5a, we can observe that the values are distributed throughout the entire range in an increasing fashion. This implies that increasing the MAD loosens the alignment restrictions, resulting in a higher percentage of matched snapshots.
- $\mathbf{P_{op}}$ coefficients are really low for all samples, affecting around 3.6% for each unit we increase the penalty, which means that it barely affects the percentage of aligned snapshots.
- $\mathbf{P_{ex}}$ coefficients are high considering all samples, as a change of 1 unit in P_{ex} would increase the percentage of aligned snapshots by 15.9%. However, when we only consider the aforementioned segment in which we obtain the optimal results, this coefficient is reduced to 7.9%. This indicates that modifying the value of P_{ex} within the appropriate range of values [-1, 0) has little impact on the percentage of aligned snapshots.

The conclusion is that we can achieve satisfactory results for the incubator example with $\mathbf{P_{ex}}$ **values between [-1.0, 0)**, regardless of the specific variations within that range.



(a) Ordered by % of matched snapshots.

(b) Ordered by the number of gaps.

Fig. 5: Analysis of alignment statistics for simple gap for the 4-P Model.

TABLE II: Analysis of the influence of affine gap penalty on the percentage of aligned snapshots.

Model	R-squared	F-statistic	Coef. MAD	P-value MAD	Coef. P_{op}	P-value P_{op}	Coef. P_{ex}	P-value P_{ex}
Affine 2-P (segment)	0.669	144.0	37.844 ± 1.915	0.000	0.728 ± 0.072	0.000	2.856 ± 0.326	0.000
Affine 2-P (all)	0.871	1876.0	87.953 ± 5.959	0.000	4.031 ± 0.204	0.000	25.003 ± 0.353	0.000
Affine 4-P (segment)	0.637	23.94	32.047 ± 4.930	0.000	1.351 ± 0.176	0.000	7.959 ± 1.564	0.000
Affine 4-P (all)	0.962	6895.0	82.027 ± 2.084	0.000	3.661 ± 0.071	0.000	15.985 ± 0.123	0.000

2) Affine gap

I. 2-Parameters model

The analysis for the Affine Gap approach is similar to that performed for the simple approach. In Figures 6a and 6b, we have the same statistical analysis for the Affine Gap approach. The first plot displays the statistics sorted by % of matched snapshots, while the second plot sorts them by the number of gaps.

If we look at Figure 6a, the results are similar to those of the Simple Gap. MAD values in the segment with the highest percentage of alignments are evenly distributed throughout the range. The same happens for P_{op}, which shows that we can get satisfactory alignments for any value within the range of [-3, -0.5]. As for P_{ex}, the optimal values are obtained within the range [-0.5, 0). The algorithm requires the gap costs not to be too high in order to include gaps to characterize missing behavior and obtain relevant alignments. In this optimal range, the Euclidean Average distance reaches its lowest value. The Fréchet distance is influenced by the MAD, having similar fluctuations when comparing the graphics.

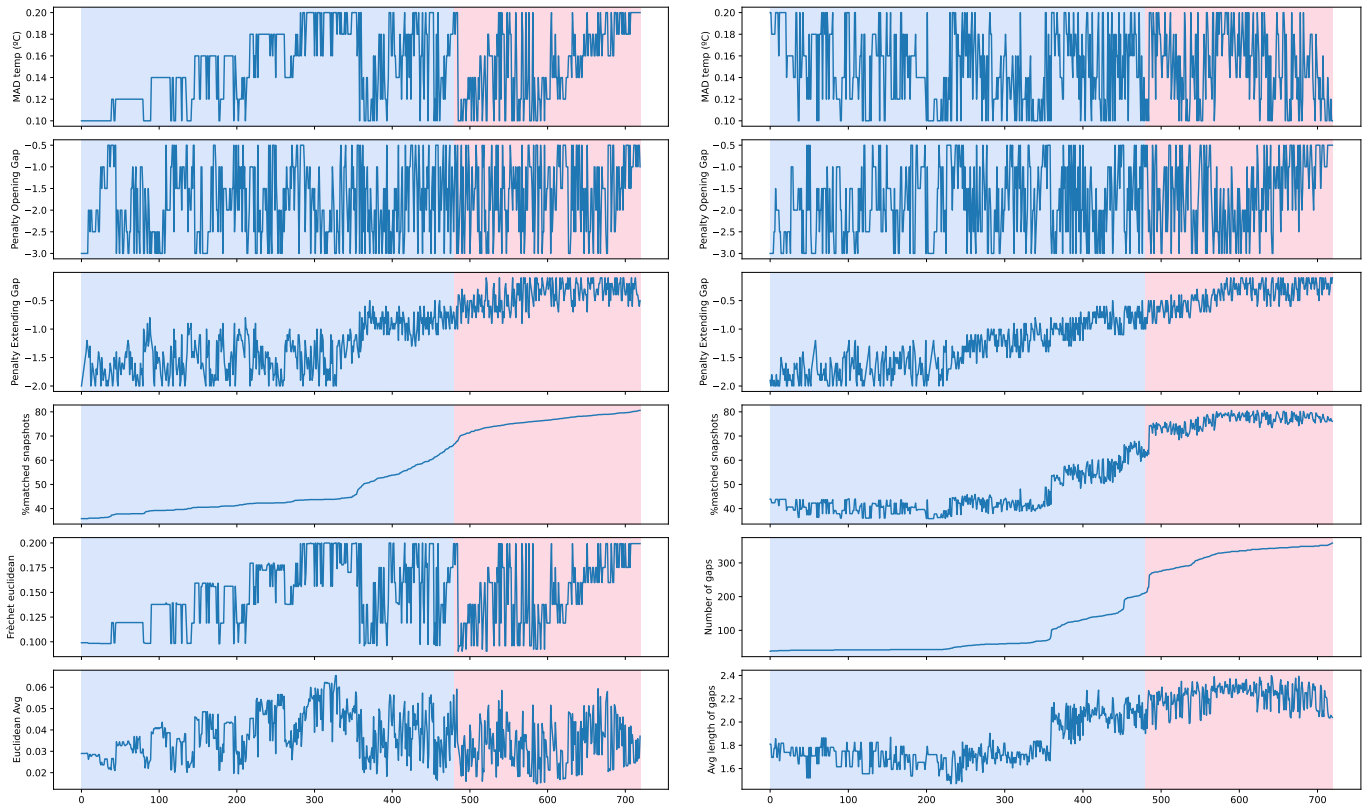
Similarly, in Figure 6b, as we reduce the cost of P_{ex}, we increase the number of gaps that the algorithm adds to the alignment, allowing flexibility in the alignment choices and improving the number of aligned snapshots. We can also observe that as we reduce the cost of the gaps, the average length of the gaps decreases, prioritizing shorter gaps over longer ones.

Regarding the analysis of the statistical significance of the data, the results are in Table II (Affine-All, Affine-Segment) and are similar to the previous ones.

- **MAD** has a relatively high coefficient: if we change one degree of the value of MAD, the % of aligned snapshot should change approximately the value of the coefficient, which is approximately 87%. However, by examining Figure 6a, we can observe that increasing the MAD loosens the alignment restrictions, resulting in a higher percentage of matched snapshots.
- $\mathbf{P_{op}}$ coefficient is low for the segment (< 1) in which its value is between [0, -0.5]. This means that a one-unit change in the P_{op} produces a change smaller than 1 in the % of matched snapshots in the segment.
- $\mathbf{P_{ex}}$ coefficient is also low for the segment in comparison to all samples with a value of 2.8 against 25. This means that the variation of P_{ex} within the range of [-0.5, 0) does not produce remarkable changes.

Therefore, the appropriate configurations for the algorithm would include a $\mathbf{P_{ex}}$ **value between [-0.5, 0) and an** $\mathbf{P_{op}}$ **value between (-3, 0).** In our approach, we typically use the combination of **-1 as** $\mathbf{P_{op}}$ **and -0.1 as** $\mathbf{P_{ex}}$, which is one of the recommendations from BLAST.

BLAST suggests that the penalty for initiating a gap should be 10 to 15 times higher than the penalty for extending it. The values for the penalties of opening and extending a gap for BLAST are obtained empirically and usually depend on the frequency scoring matrix used for the alignment [4]. However, generally, as a default value, the penalty for opening is approximately ten times higher than the cost for continuing a gap.



(a) Ordered by % of matched snapshots.

(b) Ordered by the number of gaps.

Fig. 6: Analysis of alignment statistics for affine gap for the 2-P Model.

II. 4-Parameters model

The analysis for the 4-Parameters model is similar to the one performed on the 2-Parameter model. In Figures 7a and 7b, we have the same statistical analysis for the Affine Gap approach. The first plot displays the statistics sorted by % of matched snapshots, while the second plot sorts them by the number of gaps.

If we look at Figure 7a, the results are similar to those of the previous section. MAD values in the segment with the highest percentage of alignments are evenly distributed throughout the range. The same happens for P_{op}, which shows that we can get satisfactory alignments for any value within the range of [-3, -0.5]. As for P_{ex}, the optimal values are obtained, just like in the previous case with the 2-P Model, within the range [-0.5, 0). The algorithm requires the gap costs not to be too high in order to include gaps to characterize missing behavior and obtain relevant alignments. In this optimal range, the Euclidean Average distance reaches its lowest value. The Fréchet distance is influenced by the MAD, having similar fluctuations when comparing the graphics.

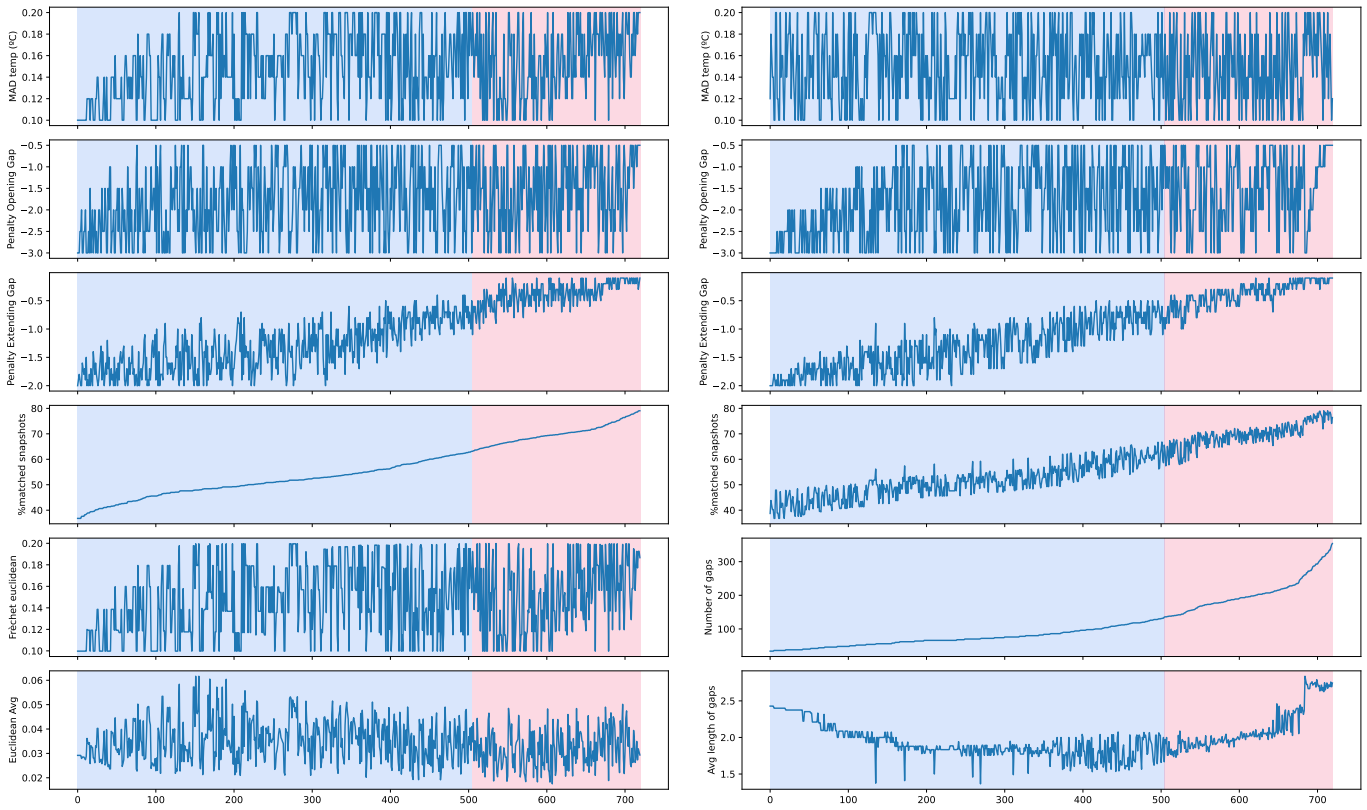
Similarly, in Figure 7b, as we reduce the cost of P_{ex}, we increase the number of gaps that the algorithm adds to the alignment, allowing flexibility in the alignment choices and improve the number of aligned snapshots. We can also observe that as we reduce the cost of the gaps, the average length of the gaps decreases, prioritizing shorter gaps over longer ones.

Regarding the analysis of the statistical significance of the data, the results are in Table I (Affine-All, Affine-Segment) and are similar to the previous ones.

- **MAD** has a relatively high coefficient: if we change one degree of the value of MAD, the % of aligned snapshot should change approximately the value of the coefficient, which is approximately 82%. However, by examining Figure 6a, we can observe that increasing the MAD loosens the alignment restrictions, resulting in a higher percentage of matched snapshots.
- $\mathbf{P_{op}}$ coefficient is low for the segment (< 2) in which its value is between [0, -0.5]. This means that a one-unit change in the P_{op} produces a change smaller than 2 in the % of matched snapshots in the segment.
- $\mathbf{P_{ex}}$ coefficient is also low for the segment in comparison to all samples with a value of 7.9 against 15.9. This means that the variation of P_{ex} within the range of [-0.5, 0) does not produce remarkable changes.

Therefore, the appropriate configurations for the algorithm would include a $\mathbf{P_{ex}}$ **value between [-0.5, 0) and an** $\mathbf{P_{op}}$ **value between (-3, 0).** In our approach, we typically use the combination of **-1 as** $\mathbf{P_{op}}$ **and -0.1 as** $\mathbf{P_{ex}}$, which is one of the recommendations from BLAST.

BLAST suggests that the penalty for initiating a gap should be 10 to 15 times higher than the penalty for extending it. The values for the penalties of opening and extending a gap for BLAST are obtained empirically and usually depend on the frequency scoring matrix used for the alignment [4]. However, generally, as a default value, the penalty for opening the score is approximately ten times higher than the cost for continuing a gap.



(a) Ordered by % of matched snapshots.

(b) Ordered by the number of gaps.

Fig. 7: Analysis of alignment statistics for the affine gap for the 4-P Model.

D. Fidelity assessment

Verifiability: The analysis performed in this section is available at https://github.com/atenearesearchgroup/fidelity-measure-for-dts/blob/main/src/evaluation/incubator/incubator_comparison_analysis.ipynb

Next, we analyze the level of fidelity we achieve when aligning the simulator trace with the trace of the same behavior in the real system. For this analysis, we use the some of the fidelity metrics we have defined in Section III of the General Concepts Technical Report [3]: the percentage of aligned snapshots, the Frèchet distance, and the average Euclidean distance (in the relevant area) between aligned snapshots.

To interpret these metrics, we need to consider what the values would be for the alignment of two identical traces. In that case, for any value of MAD:

- the **percentage of aligned points** would be 100%
- the **Frèchet distance** would be zero
- the **average Euclidean distance** would also be zero

These would be the results we would obtain for a model that had the maximum level of fidelity and was capable of accurately emulating the system. Anything that deviates from this model indicates a lower level of fidelity. We can compare different models and assess their fidelity level based on the metrics, using perfect alignment as a reference. In our work, we proposed a set of fidelity indicators to decide the degree of fidelity of a DT with respect to a PT depending on the values of the three metrics.

- **Alignment with *%MS* above 95%** ($\pm 2\%$)[1] is considered good enough, and the degree of fidelity depends on the distance between the traces.
- If **%*MS* is between 90% and 95%** ($\pm 2\%$)**,** alignment is low, but the distance metrics can be considered. The acceptable distance between the traces is application-dependent, and whether it is the Fréchet or the Euclidean distance that really matters.
- **If *%MS* below 90%** ($\pm 2\%$): traces could not be properly aligned, and therefore no faithful behavior can be expected.

The alignment algorithm applies the following configuration for all scenarios:

Parameter	Range	Increments
Maximum Acceptable Distance (MAD)	[0.2, 2.0]	0.2
Penalty opening a gap (P_{op})	-1	-
Penalty extending a gap (P_{ex})	-0.1	-

The specific and detailed guidelines on how to set the configuration values for Affine Gap are available in the previous section. In the incubator experiment, we do not consider any Low Complexity Areas. As for MAD, it was empirically established by determining where the plateau of fidelity metrics was achieved for illustrative purposes. To establish a single value in a practical example, we need to reason about the maximum distance we want to allow for aligning two snapshots. We will further develop this idea in the subsequent sections based on the data. In the following sections, we will assess the level of fidelity of both models and compare their suitability as a digital twin depending on the requirements we impose on the system.

[1]Note that we are assuming a maximum permissible error (MPE) of 2% [5] for the assessment of *%MS*, since most times thresholds are not completely accurate.

1) Heating time 3 s - Heating gap 2 s (Ht3Hg2)

To illustrate this scenario, we included alignment figures for both models with MAD values $0.2°C$ (see Figure 9) and $1.2°C$ (see Figure 10). To enhance the visualization of the alignments, note that a constant offset of $3°C$ has been added to all DT snapshots, preventing overlapping.

The results for the three fidelity metrics for different MAD values are available in Table III and a visual representation of this data in Figure 2. In this figure, we present a comparison between the two-parameter model and the four-parameter model, and we observe that the values of the three metrics are very close. The difference in the percentage of aligned snapshots is approximately $\pm 2\%$, while the two distances differ by less than 0.1 in most cases.

Next, we will analyze the results from the lowest MAD value and onwards. For the lowest MAD value, $0.2°C$, the 2-parameter model achieves 79.97% of aligned snapshots, while the 4-parameter model falls slightly behind at 78.54%, with similar distance metrics. One might logically think that since the 4-parameter model is more complex, it should better emulate the behavior of the real system. However, when we examine Figure 12a and check the alignment for different transitions, we observe fewer gaps in the case of the 2-parameter model. For instance, the two transitions occurring between timestamps 1000 and 1500 have fewer gaps in the 2-parameter model.

If we continue increasing the value of MAD, we can observe how the 4-parameter model takes the lead in terms of the percentage of aligned snapshots, surpassing the 2-parameter model by 2% or 3% starting from $0.6°C$. However, this increment comes at the cost of significantly increasing the Fréchet distance, which exceeds that of the 2-parameter model by approximately 0.2 between $0.6°C$ and $1°C$. Afterward, this difference decreases and remains at 0.1 or less.

To assess the appropriate MAD level in other case studies like the elevator scenario [6], we have looked for the point where the values reach a plateau, and the alignment statistics remain stable despite increasing the MAD. In those cases, we concluded that the suitable MAD value is 2 or 3 times the precision of the measuring device. However, in this current scenario, we haven't found such a plateau, and we observe a continuous increase in the metrics. This is typical of low-quality alignments, where as we increase the MAD, more alignments are permitted due to the previously mismatched snapshots that can be aligned at higher MAD values. Nonetheless, a tolerance of greater than $1.5°C$ is not acceptable, as it does not fit the system requirements.

In this case, to assess the fidelity level, we will consider the MAD level between 2 and 3 times the precision of the measuring device, which is $\pm 0.5°C$. Therefore, we take the range [1, 1.5]. The midpoint of this range could be 1.2. For this value, we have the alignments shown in Figure 10. Compared to the previous figure, we can observe smaller gaps between the transitions.

With this information, we can draw two conclusions:

- **No plateau is reached within the required range** for any of the models. The stakeholders consider that two snapshots with a difference above $\pm 1.5°C$ cannot be considered equivalent, according to the system requirements. The accuracy of the temperature sensor is $\pm 0.5°C$, so considering that the suitable MAD is 2 or 3 times this accuracy, we perform the analysis in the range [1, 1.5].
- Within this range of MAD, **the percentage of aligned snapshots is below 90%** ($\pm 2\%$) for both models, so we conclude that the models fail to replicate the behavior in the given conditions.

TABLE III: Fidelity results for scenario Ht3Hg2.

MAD (ºC)	2-P Model % matched	2-P Model Frèchet	2-P Model Avg. Euclidean	4-P Model % matched	4-P Model Frèchet	4-P Model Avg. Euclidean
0.2	79.9796	0.1994	0.0338	78.5495	0.1926	0.0349
0.4	82.5332	0.2543	0.0547	83.9632	0.3928	0.0646
0.6	83.7589	0.3104	0.0756	87.1297	0.5493	0.0958
0.8	85.2911	0.4311	0.1004	88.5598	0.5178	0.1154
1	86.4147	0.4941	0.1226	89.7855	0.5913	0.1358
1.2	88.049	0.6191	0.1549	90.9091	0.6713	0.159
1.4	89.1726	0.7126	0.1806	91.8284	0.7346	0.1789
1.6	90.5005	0.7751	0.2146	92.4413	0.7796	0.1945
1.8	91.522	0.9001	0.2411	93.3606	0.8252	0.2196
2	92.5434	0.9626	0.272	94.1777	0.8846	0.2439

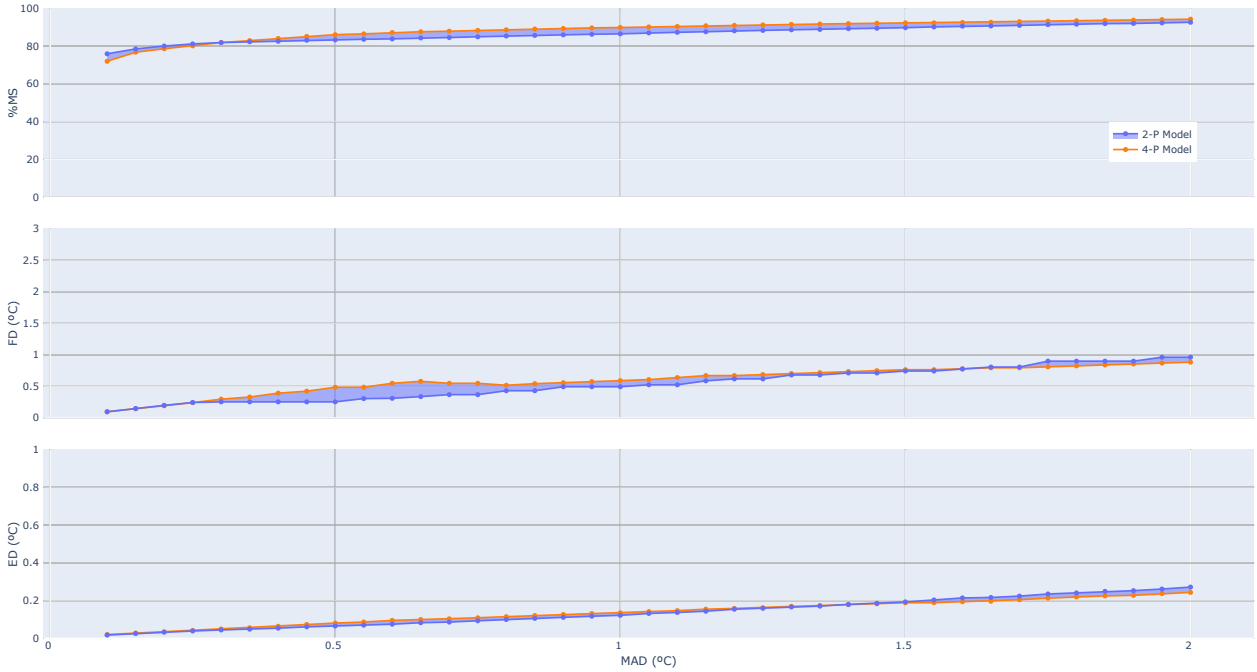
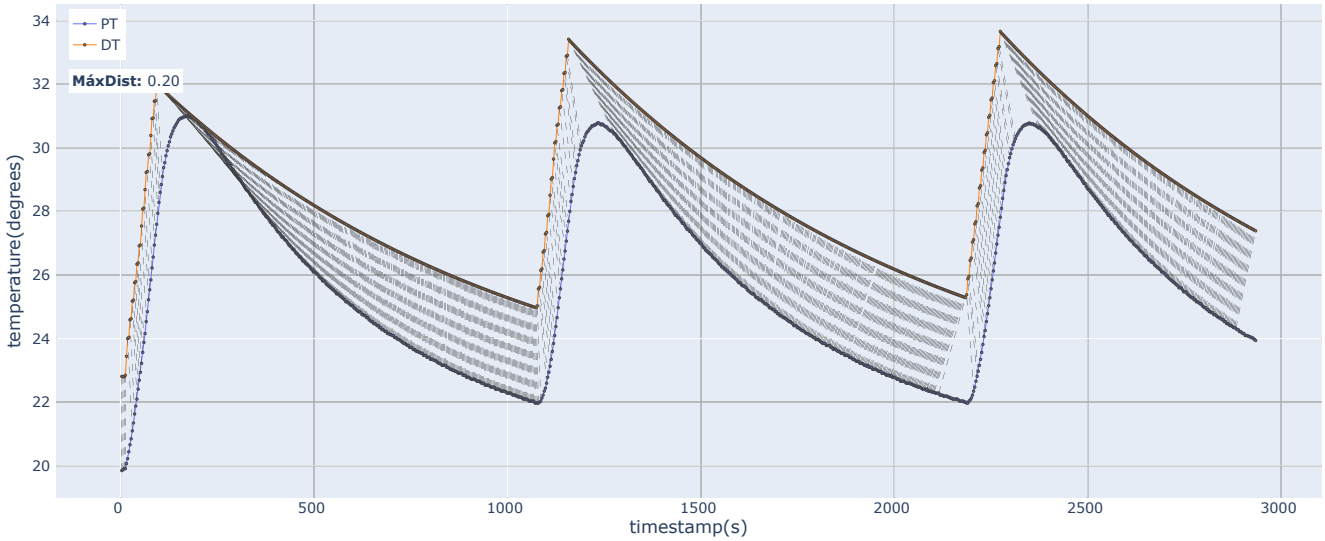
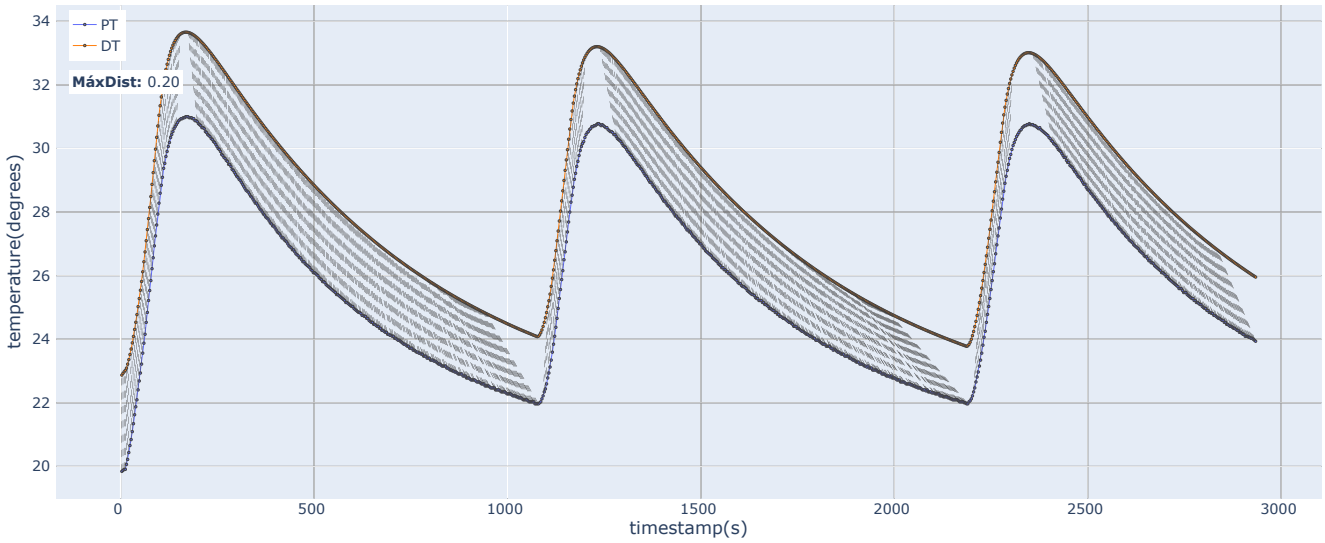


Fig. 8: Statistical comparison of the 2-P Model against the 4-P Model for scenario Ht3Hg2.

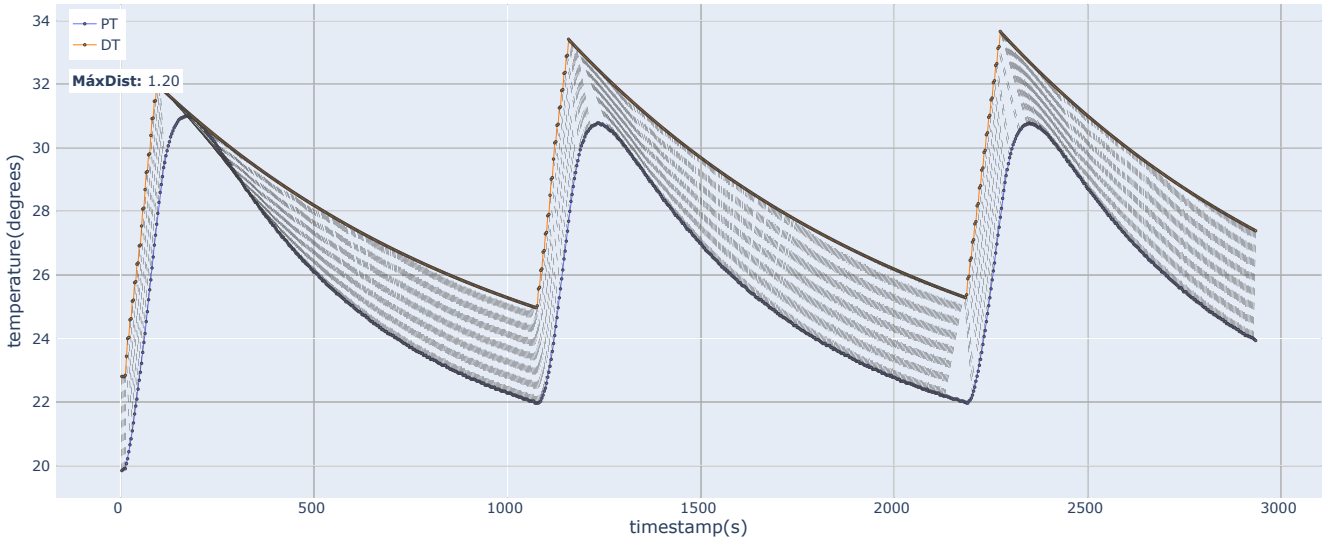


(a) 2-P Model

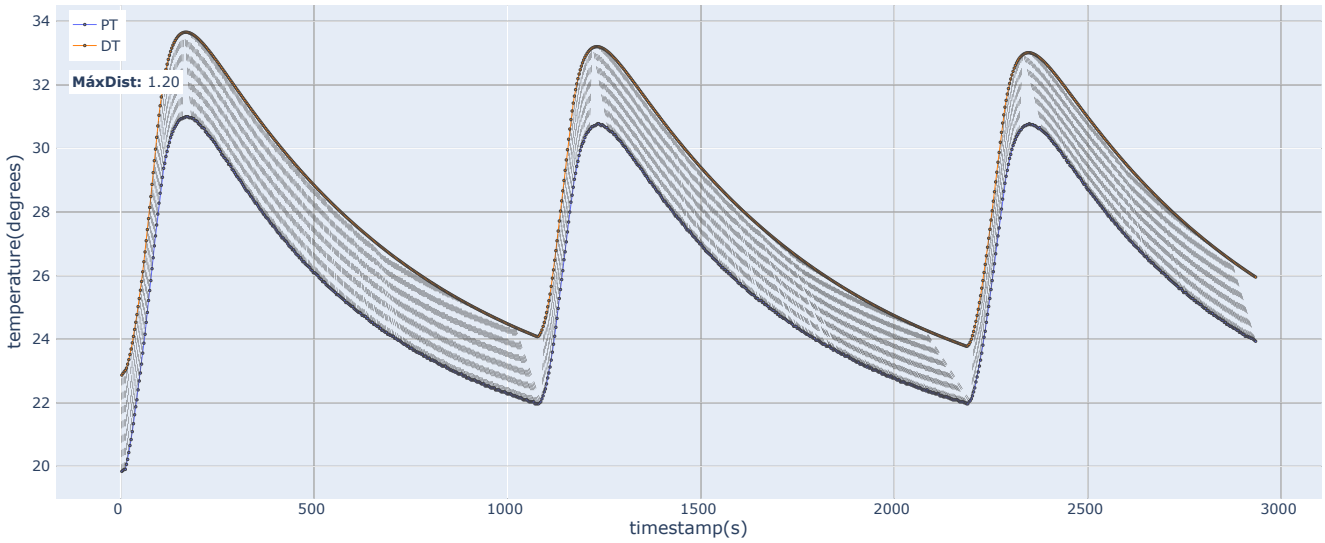


(b) 4-P Model

Fig. 9: Alignments for scenario Ht3Hg2 with MAD 0.2.



(a) 2-P Model



(b) 4-P Model

Fig. 10: Alignments for scenario Ht3Hg2 with MAD 1.2.

2) Heating time 30 s - Heating gap 20 s (Ht30Hg20)

To illustrate this scenario, we included alignment figures for both models with MAD values $0.2°C$ (see Figure 12) and $1.2°C$ (see Figure 13). To enhance the visualization of the alignments, note that a constant offset of $3°C$ has been added to all DT snapshots, preventing overlapping.

The results for the three fidelity metrics for different MAD values are available in Table IV and a visual representation of this data in Figure 3. In this figure, we present a comparison between the two-parameter model and the four-parameter model, and we observe that the statistics of the percentage of aligned snapshots are much higher for the 4-parameter model, surpassing the two-parameter model by more than 10%. Furthermore, the distance statistics are much worse for the two-parameter model, reaching up to $0.4°C$, while for the four-parameter model, this value remains around $0.1°C$.

Next, we will analyze the results from the lowest MAD value and onwards. For the lowest MAD value, $0.2°C$, the 2-parameter model achieves 54.13% of aligned snapshots, while the 4-parameter model already has 89.65%, with similar distance metrics. Estos resultados contrastan enormemente con los que hemos obtenido en el análisis del escenario anterior ya que siendo los mismos models, la capacidad de emular el comportamiento es mucho mejor en este caso para un modelo y mucho peor para el otro. Esto tiene que ver con que en este escenario los períodos de calentamiento y enfriamiento son mucho más largos. Recordemos, en este escenario, el controlador detecta si se ha alcanzado la temperatura objetivo, si no es así, calienta durante 30 s y después apaga el calentador durante 20 s. Una vez pasa este tiempo, vuelve a comprobar si se ha alcanzado la temperatura objetivo y así sucesivamente.

If we analyze figure 12a, we can observe that for the 2-parameter model, the turn-off and turn-on sequence is easily noticeable, resulting in a saw blade-like pattern for the heating process. The 20-second off period is visible as a small dip in the heating process, which reduces the accuracy of the model during this interval. However, if we examine figure 12b, we can see that the turn-off is reflected as a slight loss of slope in the heating curve, which accurately simulates the behavior of the real system. Therefore, the four-parameter model provides a much better alignment for the heating processes. In this case, even the transitions are captured more precisely than in the previous scenario.

By increasing the MAD, we can observe that the 4-parameter model includes a few more snapshots to the alignment while keeping almost all the values stable and reaching a plateau. However, the two-parameter model continues to grow, but at the expense of worsening its statistics.

As stated earlier, the 4-parameter model exhibits a plateau of values around [1, 1.5]. We can observe that at these values, over 95% of the aligned snapshots have very small distances. Additionally, in Figure 13b, we can see that there are almost no gaps in the transitions, and the trajectory is almost entirely aligned.

On the other hand, when we examine the statistics of the 2-parameter model, we do not observe a plateau. If we assess the metrics between the values of 1 and 1.5, we can see that the figures stay below 75%. In Figure 13a, we can observe the model's ability to align some snapshots in the heating process, but it still maintains a large number of gaps throughout the entire trace.

With this information, we can draw different conclusions for each model:

2-parameters Model

- **No plateau is reached within the required range** for any of the models. The stakeholders consider it unacceptable a MAD above $\pm 1.5°C$. The accuracy of the temperature sensor is $\pm 0.5°C$, so considering that the suitable MAD is 2 or 3 times this accuracy, we perform the analysis in the range [1, 1.5].
- Within this range of MAD, **the percentage of aligned snapshots is below 90% (**$\pm 2\%$**)** so we conclude that the 2-parameters model fails to replicate the behavior in the given conditions.

4-parameters Model

- **Plateau is reached within the required range** at a MAD, which is 2 or 3 times the accuracy of the temperature sensor. The accuracy of the temperature sensor is $\pm 0.5°C$, so we perform the analysis in the range [1, 1.5].
- Within this range of MAD, **the percentage of aligned snapshots is above 95% (**$\pm 2\%$**)** f so we conclude that the 4-parameters model faithfully replicates the behavior in the given conditions.

TABLE IV: Fidelity results for scenario Ht30Hg20.

MAD (ºC)	2-P Model % matched	2-P Model Frèchet	2-P Model Avg. Euclidean	4-P Model % matched	4-P Model Frèchet	4-P Model Avg. Euclidean
0.2	54.1481	0.1796	0.0428	89.6521	0.1956	0.0352
0.4	57.8055	0.3901	0.0833	93.8448	0.347	0.0585
0.6	62.9795	0.5982	0.13	95.7181	0.5645	0.0776
0.8	67.4398	0.7778	0.1729	96.521	0.658	0.089
1	72.2569	0.8907	0.2281	97.0562	0.658	0.1006
1.2	75.3791	0.9124	0.2801	97.5914	0.658	0.114
1.4	77.2525	0.9747	0.3199	98.0375	0.658	0.1274
1.6	80.1963	1.0601	0.3741	98.2159	0.658	0.1343
1.8	81.8912	1.2247	0.4228	98.3943	0.658	0.141
2	83.7645	1.4437	0.4748	98.6619	0.658	0.1528

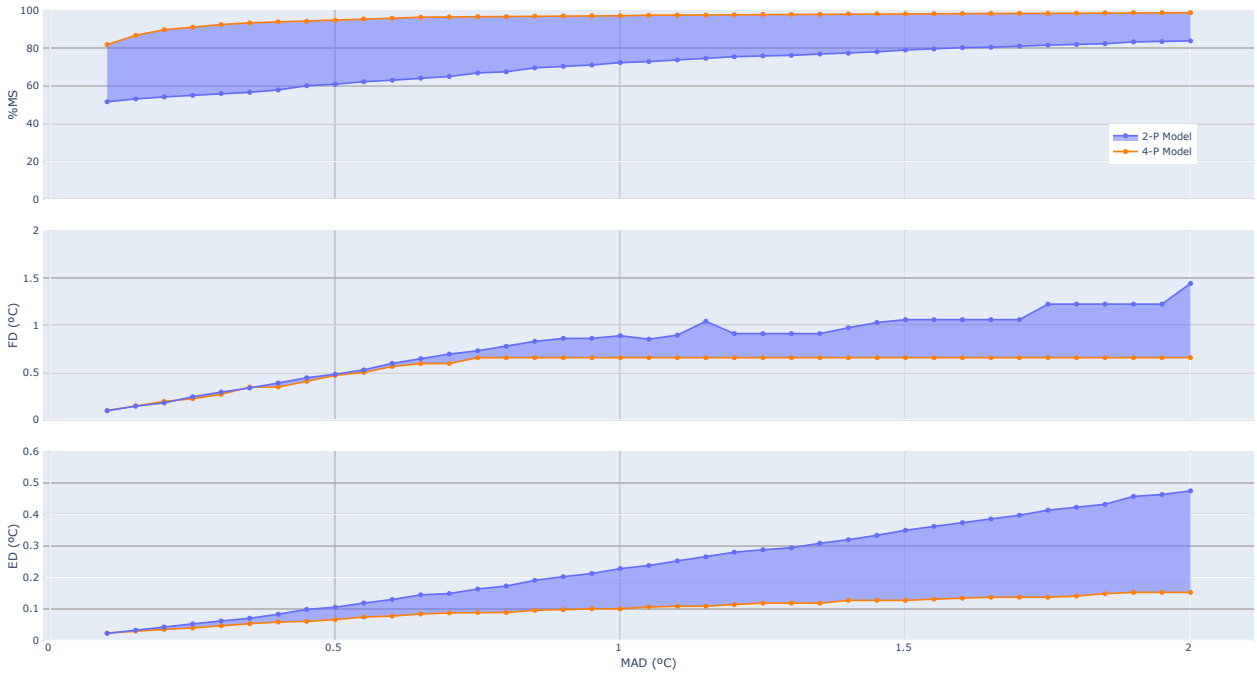
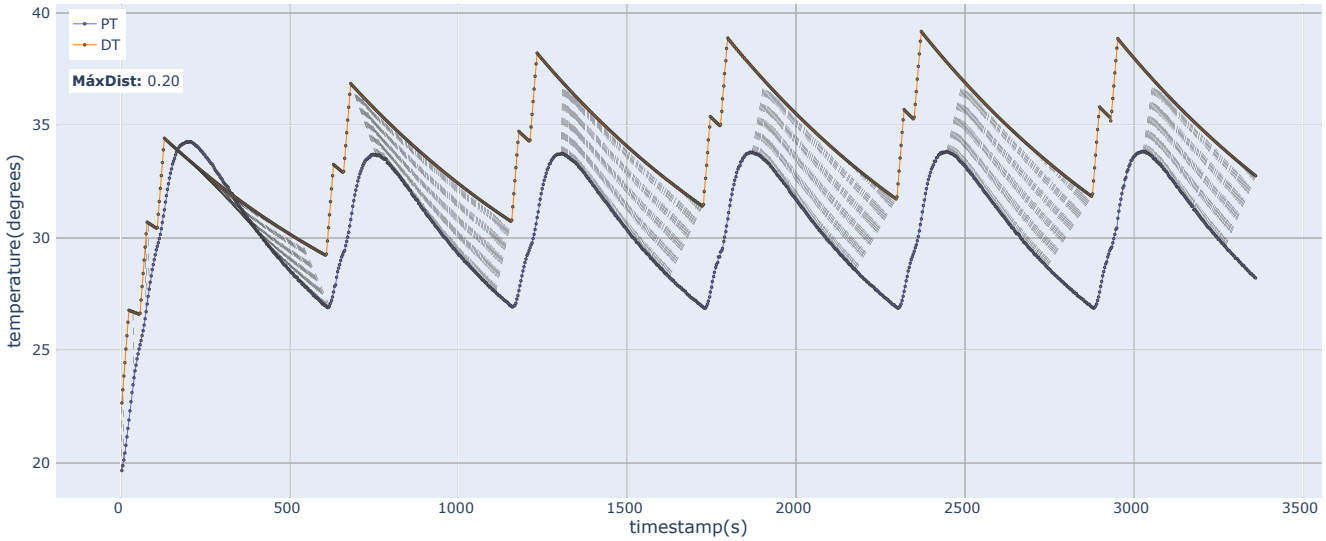
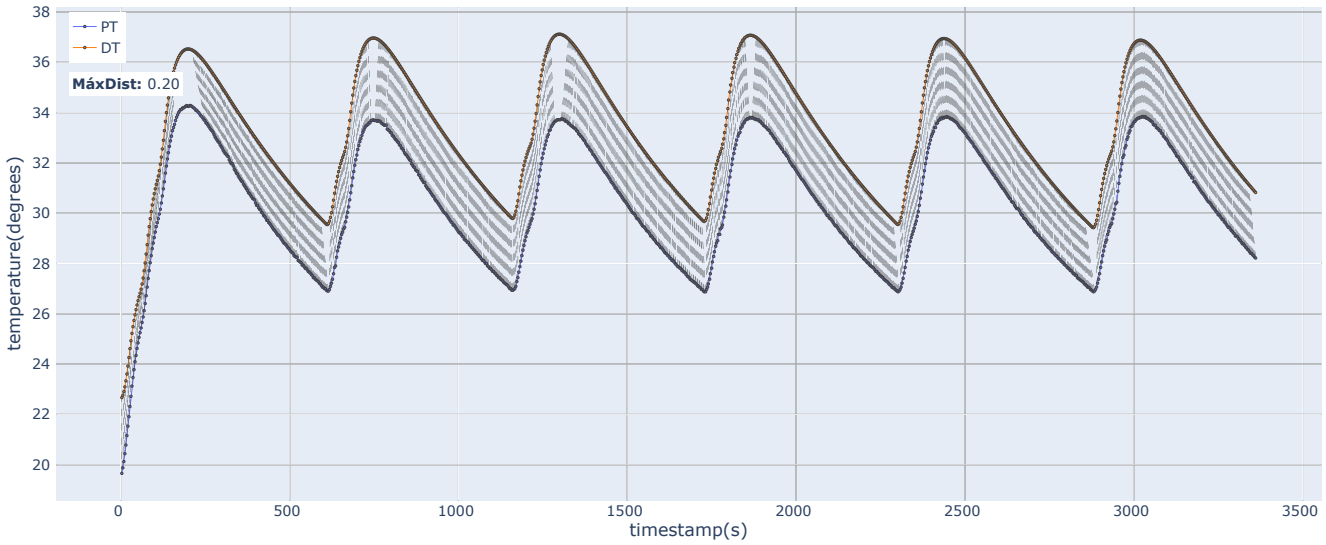


Fig. 11: Statistical comparison of the 2-P Model against the 4-P Model for scenario Ht30Hg20.

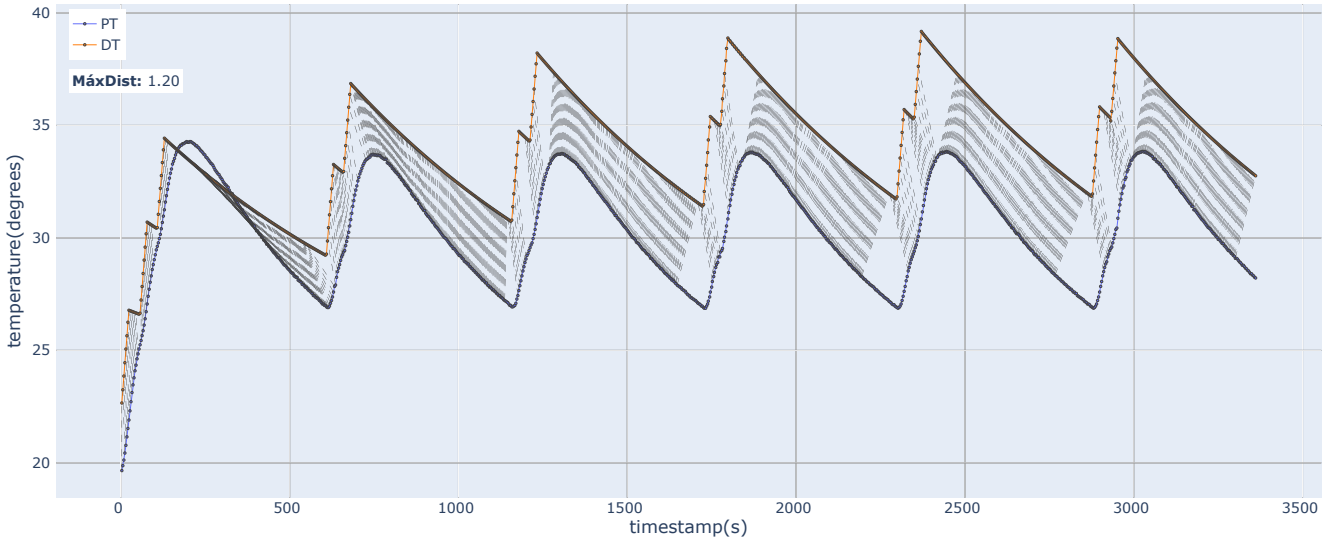


(a) 2-P Model

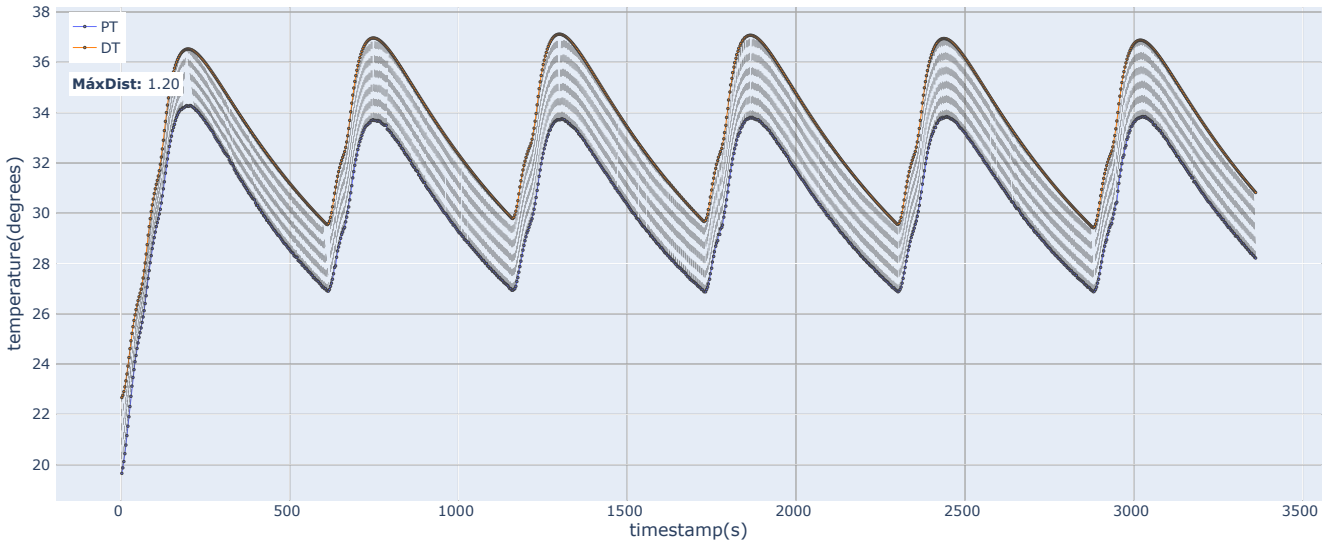


(b) 4-P Model

Fig. 12: Alignments for scenario Ht30Hg20 with MAD 0.2.



(a) 2-P Model



(b) 4-P Model

Fig. 13: Alignments for scenario Ht30Hg20 with MAD 1.2.

E. Synthetic scenarios analysis

Verifiability: The alignments performed in this section are available at https://github.com/atenearesearchgroup/fidelity-measure-for-dts/blob/main/src/resources/output/incubator/synthetic_example

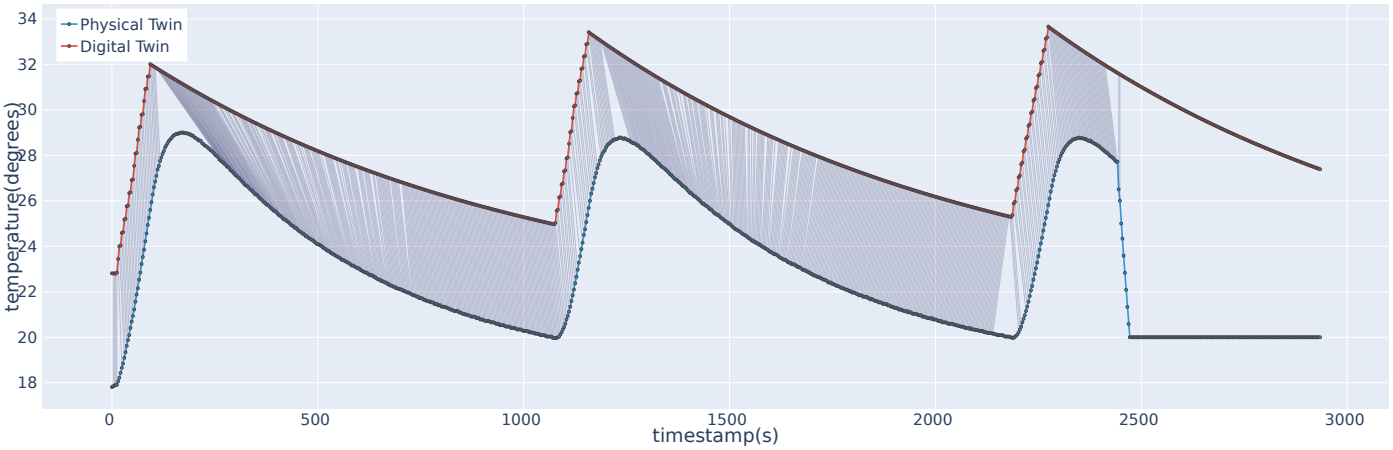
In order to test the algorithm's capabilities, we created one more synthetic scenario that included anomalies in the PT's behavior. This scenario is designed to demonstrate the algorithm's ability to identify and highlight inconsistencies between the behaviors of both systems. For this, we will use a MAD of $1.5m/s^2$, the recommended value for trace alignment in this system.

In particular, we made changes to the PT trace of the Ht3Hg2 scenario to simulate opening the incubator lid during its execution. As a result, the temperature drops to room temperature since the incubator can no longer heat a space larger than the intended box. We attempted to align this trace with both the 4-P and 2-P models, and the resulting figures can be found in Figure 14.

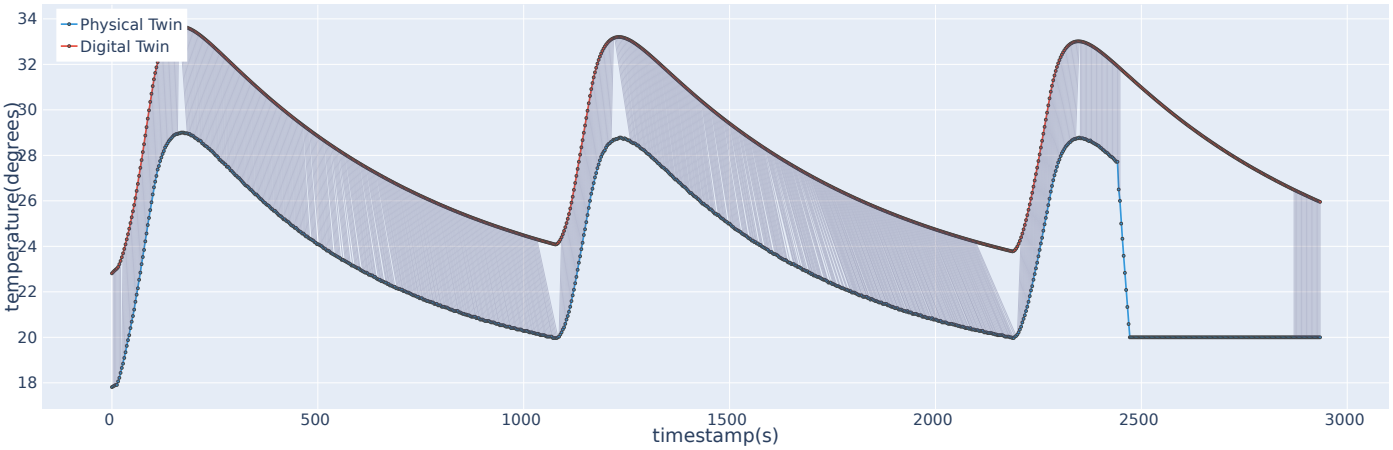
The comparison of fidelity metrics with respect to the original scenario is presented in Table V. In both cases, the percentage of matched snapshots drops, and the drop is more significant in the case of the 4-Parameter model, going from 98% to 79%. The distances are not heavily affected, and only the Frèchet distance increases by 0.8 degrees in the case of the 4-parameter model.

TABLE V: Fidelity results for scenario Ht30Hg20-open lid.

	2-P Model			4-P Model		
Scenario	% matched	Frèchet	Avg. Euclidean	% matched	Frèchet	Avg. Euclidean
Original	77.2525	0.9747	0.3199	98.0375	0.658	0.1274
Open lid	73.9530	0.7435	0.2019	79.8774	1.482	0.2494



(a) 2-P Model



(b) 4-P Model

Fig. 14: Alignments for scenario Ht3Hg2 with MAD 1.5, opening the lid in the PT trace.
The figure includes a $5\ C$ *difference in the PT's trace to improve visualization.*

II Acknowledgments

We would like to express our gratitude to Claudio Gomes and his team at Aarhus University for providing us with the data from their experiments with their digital twin of the incubator. Thanks to this data, we were able to conduct an analysis of a multi-fidelity digital twin and perform a comparative study in two very interesting scenarios.

References

[1] H. Feng, C. Gomes, C. Thule, K. Lausdahl, M. Sandberg, and P. G. Larsen, "The incubator case study for digital twin engineering," 2021.

[2] "Example digital twin: The incubator," 2023. [Online]. Available: https://github.com/INTO-CPS-Association/example_digital-twin_incubator

[3] P. Muñoz, J. Troya, M. Wimmer, and A. Vallecillo, "Using trace alignments for measuring the fidelity of a physical and a digital twin: General concepts," 2023. [Online]. Available: https://github.com/atenearesearchgroup/fidelity-measure-for-dts/blob/main/docs/Technical_Report_General_Concepts.pdf

[4] I. Korf, M. Yandell, and J. A. Bedell, *BLAST - an essential guide to the basic local alignment search tool*. O'Reilly, 2003.

[5] D. A. Snow, Ed., *Plant Engineer's Reference Book*, 2nd ed. Elsevier, 2003.

[6] P. Muñoz, J. Troya, M. Wimmer, and A. Vallecillo, "Using trace alignments for measuring the fidelity of a physical and a digital twin: Elevator technical report," 2023. [Online]. Available: https://github.com/atenearesearchgroup/fidelity-measure-for-dts/blob/main/docs/Technical_Report_Elevator.pdf