Understanding the Impact of Temporal Aggregation on Uncertainty in Quality Indicator Prediction for Industrial Processes

Thanos Kontogiannis[1], Wanda Melfo[2], Dimitrios Zarouchas[3], and Nick Eleftheroglou[4]

[1,4] *Intelligent System Prognostics Group, Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS, Delft, The Netherlands*
a.kontogiannis@tudelft.nl
n.eleftheroglou@tudelft.nl

[1,3] *Center of Excellence in AI for structures, Prognostics & Health Management, Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS, Delft, The Netherlands*
d.zarouchas@tudelft.nl

[2] *Research and Development, Tata Steel Europe, IJmuiden, 1970 CA, The Netherlands*
wanda.melfo@tatasteeleurope.com

ABSTRACT

Industrial predictive datasets often rely on coarse synchronized targets obtained by aggregating serial measurements over predefined segments. Although such aggregation is usually imposed by storage constraints and synchronization requirements, the resulting labels are commonly treated as deterministic in downstream modeling. This can be misleading when the underlying process is serially dependent, because aggregation then injects uncertainty at the label level before any model is trained and directly affects the performance that can realistically be achieved. This work examines that effect using segment-level coiling temperature prediction in hot strip steel manufacturing as a real-world example, where meter-level coiling temperature measurements are synchronized to tracked material segments and averaged to form prediction targets. A serial, dependence-aware, and deployable formulation is introduced to quantify the uncertainty associated with these aggregated targets and propagate it to the downstream predictive task. Results show that serial dependence persists in the meter-level coiling temperature series even after downsampling, and that slower sampling increases the uncertainty associated with the aggregated labels. The estimated aggregation uncertainty is further shown to be of the same order as the error achieved by an extensively optimized downstream predictor, indicating that a non-negligible portion of the apparent prediction error is attributable to noise introduced during target construction rather than to deficiencies of the predictive model alone. The findings highlight that aggregation design should not be treated as an innocent preprocessing choice. Instead, it should be considered as a factor that directly shapes attainable predictive performance, with sampling frequency emerging as an actionable lever for improving the performance ceiling of industrial datasets constructed from serial measurements.

1. INTRODUCTION

Industrial production environments routinely balance information fidelity against operational cost, which has led to the widespread use of slow sampling strategies for process monitoring and quality assurance. Rather than retaining high-frequency sensor streams in full, raw measurements are commonly downsampled and converted into window-based indicators such as averages, extrema, or exceedance rates that can be stored, synchronized, and used operationally. In many cases, these aggregated quantities are then treated as precise observations in downstream analysis, prediction, and decision-making, and the resulting models are reported almost exclusively through deterministic point metrics.

This work challenges that treatment, for the case where targets are formed by aggregating serial industrial data. When the underlying signal within an aggregation window is noisy and temporally dependent, the resulting target is not an exact observation but a realization of a local average with label-side uncertainty induced by the sampling and aggregation design itself. This distinction matters because the uncertainty is injected upstream, before any predictive model is trained, and therefore affects the level of performance that can realistically be achieved downstream. In that sense, aggregation uncer-

tainty is not simply a nuisance term, but a design-induced ceiling that should be quantified before increasingly complex models are pursued.

This perspective is largely absent from representative industrial prediction and control studies, which typically focus on improving point accuracy on aggregated or segment-level targets. In the hot-strip-mill context, coiling-temperature prediction and control have been addressed through control-oriented thermal models, analytical formulations, regression-based approaches, and neural-network-based methods, with performance discussed primarily in terms of prediction or control error rather than uncertainty attached to the aggregated target itself (Yuasa et al., 1990; Timm, Weinzierl, Leipertz, Zieger, & Zouhar, 2002; Zheng, Li, & Li, 2013; Wu, Sun, Peng, & Zhang, 2023; H. B. Xie, Jiang, Liu, Wang, & Tieu, 2006). Similar deterministic reporting is common in other segment-level industrial prediction problems, where synchronized or windowed targets are assessed through RMSE, R^2, or related point metrics (Mohanty, Banerjee, Santara, Kundu, & Mitra, 2021; Q. Xie et al., 2021; Deng, Sun, Peng, Hu, & Zhang, 2019).

At the same time, uncertainty quantification is widely studied in machine learning and prognostics, with most approaches focusing on uncertainty in the predictions produced by a model. Bayesian methods, ensembles, and conformal calibration are typically applied after the target variable has already been defined, and therefore treat the observed label as the reference quantity of interest (Hüllermeier & Waegeman, 2021; Gawlikowski et al., 2023; Fontana, Zeni, & Vantini, 2023). In the present case, however, uncertainty is introduced earlier, during target construction itself, because the value used for prediction is obtained by aggregating a noisy, serially dependent measurement process. The main question is therefore not how uncertain the predictor is around a fixed target, but how uncertain the target is before any model is trained. For variance calculation in industrial production data of this kind, care must therefore be taken not to assume independence of the aggregated observations, since doing so leads to an underestimation of the uncertainty attached to the target itself.

These issues arise directly in the real-world steel manufacturing use case considered in this study, where coiling temperature (CT) is a critical process variable linking upstream hot rolling conditions with downstream material properties and is therefore actively controlled in production. For it to be adequately controlled, it needs to be predicted before it is measured. While CT is measured with an infrared pyrometer at 1 m intervals along the strip, upstream process parameters cannot be matched directly by time because the material elongates and travels at varying velocities across the hot strip mill. To preserve material correspondence, the material is divided into tracked segments, and the multiple CT measurements within each segment are averaged into a single synchronized target value. This use case also reflects a broader class of industrial settings in which the target used for prediction is itself the result of a data-handling step that aggregates a serial measurement into synchronized windows. If the original series is temporally dependent, averaging does not eliminate uncertainty as quickly as would be expected under independence, because neighboring samples carry overlapping information. In that case, the target is better viewed as a noisy realization of a local average rather than as an exact deterministic value.

Building on this use case, the paper discusses the need to quantify aggregation-induced uncertainty when coarse synchronized segments are used as predictive targets. Segment-level CT prediction is used as an industrial case study. It is shown that serial dependence persists even after downsampling and that the resulting dependence-aware aggregation variance departs from naive $1/n$ scaling. It is further shown that slower sampling increases the uncertainty associated with the target itself. The purpose is not only to demonstrate this statistical effect, but to show that it has a direct impact on the achievable performance of downstream predictive models. Once quantified, this uncertainty provides a more realistic interpretation of predictive error and, importantly, an actionable handle for improving the performance ceiling of the dataset at hand through changes in the sampling design rather than through model complexity alone. The remainder of the paper is organized as follows. Section 2 introduces the problem setting and dataset. Section 3 presents the serial dependence-aware formulation used to quantify aggregation uncertainty. Section 4 reports the empirical results, including the sampling stress analysis and the link to predictive error. Section 5 concludes with implications, limitations, and possible extensions to additional label-side uncertainty sources.

2. PROBLEM STATEMENT AND TARGET CONSTRUCTION

In the hot strip mill, coiling temperature (CT) is a critical process variable because it reflects the thermal history of the strip and is directly linked to downstream material properties and process quality. It is the temperature measured at the end of the milling process, after the cooling procedure (Run-Out Table - ROT) has been completed. In the present work, CT is considered an indirect indicator of surface quality. Previous analysis of the same production context has shown that major CT deviations and rapid local fluctuations can be associated with severe surface defects, particularly when oxide layers become unstable and chip off during processing. This interpretation is physically plausible because infrared pyrometer measurements are strongly affected by surface emissivity, and the local exposure of bare steel after oxide chipping alters the measured temperature response. At the same time, direct measurement of oxide layer thickness during production is not feasible in continuous hot strip milling, since it requires interrupting the process and freezing the oxide layer for ex situ inspection (Min, Kim, Kim, & Lee, 2012). For this reason,

CT provides a practical and informative proxy for predicting surface defects (Kontogiannis, Melfo, Eleftheroglou, & Zarouchas, 2024; Kontogiannis, Zarouchas, & Eleftheroglou, 2026). This is achieved by building a predictor utilizing the process parameters directly available before the ROT to predict the CT and thus the surface defects.

The CT is measured at a high frequency corresponding to one measurement every few centimeters of the steel strip. Due to resource constraints, one measurement per meter of material is kept. Even so, the dataset built for predicting the CT is not used directly at this resolution. Two practical constraints impose aggregation. First, storage and handling requirements become substantial at the production scale when high-frequency measurements are continuously retained over multiple years of operation. Second, and more importantly in the present case, upstream process parameters cannot be synchronized with CT using timestamps alone. As the strip passes through the hot strip mill, it elongates and travels at different velocities across the different stages of the process. Consequently, measurements recorded at the same time do not necessarily refer to the same material location.

To preserve material correspondence, the slab is divided early in the process into tracked segments that follow the same material throughout production. CT measurements and upstream process variables are then aligned at the segment level. Since multiple meter-level CT readings fall within each tracked segment at the coiling station, the target used for its prediction is inherently aggregated: a single segment-level value is constructed by averaging the available CT measurements within the synchronized segment. The aggregation step is therefore not an arbitrary preprocessing choice, but a consequence of real deployment and synchronization requirements (Kontogiannis et al., 2026).

Let $T_{c,m}$ denote the meter-level coiling temperature measured at position t along coil c, sampled at spatial interval $\Delta s = 1\,\mathrm{m}$, and let $T^{\star}_{t,m}$ denote the corresponding goal coiling temperature profile used in production. Rather than modeling the raw temperature directly, the predictive target is defined as the meter-level CT deviation normalized by the goal CT, following the implementation adopted in the companion CT prediction study (Kontogiannis et al., 2026):

$$y_{t,m} = \frac{T_{t,m} - T^{\star}_{t,m}}{T^{\star}_{t,m}}. \tag{1}$$

Let segment j of coil c correspond to the index set $\mathcal{I}_{c,j}$, containing n_j meter-level samples. The segment-level predictive target is then defined as the average normalized CT deviation within the synchronized segment,

$$Y_{c,j} = \frac{1}{n_j} \sum_{t \in \mathcal{I}_{c,j}} y_{c,t}. \tag{2}$$

This is the quantity used as the prediction label in the downstream learning task, and it is the scale on which all results are reported in the present paper.

A distinction must be made between the data-analysis stage and the predictive task itself. In the training part of the dataset, both the meter-level CT series and the derived segment-level CT targets are available. This makes it possible to study the serial dependence of the meter-level signal and to estimate the aggregation uncertainty associated with the segment-level target. In contrast, for the predictive task, the available inputs consist only of synchronized segment-level process features, while CT is not available and must be predicted at the segment level. Thereby, the target used for prediction is the aggregated segment-level quantity $Y_{c,k}$, whereas the dependence analysis is performed on the normalized meter-level CT deviation series available in the training data. In this way, the aggregation uncertainty is estimated from the same underlying signal used to construct the segment-level targets, while the downstream predictive model operates only on synchronized segment-level inputs. The uncertainty quantified in this work, therefore, pertains to the construction of the target itself, rather than to the model predictions.

3. METHODS

The methodology of the present work is focused on quantifying the uncertainty introduced when meter-level coiling temperature (CT) measurements are aggregated into synchronized segment-level prediction targets. The meter-level CT measurements describe the temperature of a continuous material that has undergone a series of forming and cooling processes, all of which affect its final dimensions and properties and are directly linked to the resulting CT. It is therefore expected that these measurements exhibit serial dependence and cannot be treated as i.i.d. When handled as an ordered series, this dependence is reflected as temporal dependence, and thus the autocorrelation structure must be taken into account when calculating the variance of each aggregated segment. In the following, the methodology used to estimate the variance injected into the dataset by the aggregation of an autocorrelated series is presented. Next, the way in which this estimation can be made deployable in an online setting for unseen data is described. Finally, a simulated downsampling scheme is introduced to illustrate the effect of sampling frequency on the injected label noise and its implications for the achievable performance of downstream prediction tasks.

3.1. Aggregation variance under serial dependence and stationarity assumption

The within-coil meter-level process has autocovariance $\gamma(h)$, the variance of a finite window mean of length n is

$$\mathrm{Var}(\bar{y}_n) = \frac{1}{n^2} \sum_{s=1}^{n} \sum_{t=1}^{n} \gamma(|t-s|), \tag{3}$$

which can be written as

$$\mathrm{Var}(\bar{y}_n) = \frac{\gamma(0)}{n} \left[1 + 2 \sum_{h=1}^{n-1} \left(1 - \frac{h}{n} \right) \rho(h) \right], \tag{4}$$

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{5}$$

This expression is the key link between serial dependence and aggregation uncertainty. Under independence, $\rho(h) = 0$ for all $h > 0$, and the usual $1/n$ variance scaling is recovered. Under positive short-range dependence, the variance of the segment mean decreases more slowly, so the aggregated target is less precise than implied by naive i.i.d. scaling.

To use this formulation in a deployable way, the meter-level process is treated as second-order stationary, so that its autocovariance depends on lag rather than on absolute coil position. Under this assumption, the variance of a segment average is determined by the common lag-dependent autocovariance structure together with the segment length. This assumption is adopted here as a practical necessity. Since coils have different total lengths and the final coil length is not available as a reliable online conditioning variable at prediction time, the uncertainty map cannot be indexed by normalized position along the coil. A single lag-indexed dependence structure is therefore used so that aggregation variance can be assigned from segment geometry alone. This is also a reasonable approximation in the present setting because the strongly atypical head and tail regions, which exhibit clearly different behavior, are excluded from the predictive dataset. If more granular online information becomes available, the same framework can be extended to region-specific or regime-conditioned autocovariance models.

3.2. Train-only estimation of the dependence structure

Although the finite-n variance expression in Eq. (4) could be evaluated directly from the estimated autocovariance sequence, the aggregation variance is instead computed using a Bartlett-weighted HAC estimator (Newey & West, 1987). The reason is that direct substitution of the empirical $\hat{\gamma}(h)$ into the finite-n expression requires summing lag contributions up to $h = n - 1$, while the estimated autocovariances at larger lags are based on fewer valid pairs and are therefore substantially noisier. As a result, the direct plug-in estimator becomes sensitive to irregular high-lag estimates and may lead to unstable variance values for finite segments. The HAC formulation addresses this by truncating the lag sum and downweighting larger lags, thereby preserving the short-range dependence structure that is most relevant for aggregated industrial targets, while reducing the influence of noisy long-lag autocovariance estimates (Andrews, 1991). In production datasets such as the present one, local persistence is expected due to process inertia, thermal continuity, and sequential material transport, whereas long-lag dependence is both less directly relevant to segment averaging and more difficult to estimate reliably from finite data.

Accordingly, for a segment j containing n_j meter-level samples, the aggregation variance is estimated as

$$\widehat{v}_{\mathrm{agg}}(n_j) = \frac{1}{n_j} \left[\widehat{\gamma}(0) + 2 \sum_{h=1}^{b} w_h \, \widehat{\gamma}(h) \right], \tag{6}$$

$$w_h = 1 - \frac{h}{b+1}, \tag{7}$$

where b is the truncation bandwidth and w_h are Bartlett weights. This corresponds to the Newey-West estimate of the variance of an equal-weight mean under serial dependence (Newey & West, 1987). The bandwidth controls how many lags are retained in the variance estimate: larger values include more of the dependence structure but also more estimation noise, whereas smaller values produce a more stable but more aggressively truncated estimate (Andrews, 1991). When no bandwidth is fixed a priori, it is selected automatically using the Newey-West rule-of-thumb:

$$b = \left\lfloor 4 \left(\frac{n}{100} \right)^{2/9} \right\rfloor, \tag{8}$$

with the additional constraint that b cannot exceed the maximum lag available from the estimated autocovariance sequence (Newey & West, 1994). In this way, the fitted train-only dependence structure is converted into a stable estimate of $\widehat{v}_{\mathrm{agg}}(n)$ that can be evaluated for any relevant segment length.

3.3. Autocorrelation diagnostic

To assess whether serial dependence is present in the meter-level CT series, the autocorrelation function (ACF) is computed on the training data. Since the aggregation-variance formulation in Section 3.1 relies on lag-dependent dependence, the ACF is used here as a diagnostic quantity to verify and visualize the extent to which consecutive meter-level observations remain correlated. For each coil, the lag-h autocorrelation is computed from the same centered series used for the autocovariance estimation, and the resulting ACF curves are then averaged across coils. In this way, the ACF provides a direct empirical view of the short-range dependence structure that

motivates the use of HAC-based aggregation variance instead of naive i.i.d. scaling.

3.4. Deployable aggregation-uncertainty map

Once the autocovariance structure $\widehat{v}_{\mathrm{agg}}(n)$ has been estimated on the training split, the aggregation variance is tabulated as a lookup map from segment length to label variance,

$$n_j \mapsto \widehat{v}_{\mathrm{agg}}(n). \tag{9}$$

For any segment j, the aggregation variance is then assigned as

$$v_{\mathrm{agg},j} = \widehat{v}_{\mathrm{agg}}(n_j), \tag{10}$$

using only the segment length n_j and the train-only dependence structure. No segment label is used at this stage. This is what makes the procedure deployable: the downstream predictive model operates only on synchronized segment-level features, yet each segment can still be associated with a label-side uncertainty estimate derived from the training meter-level CT series and the known segmentation design.

For reporting and interpretation, the corresponding standard deviation and 95% confidence-interval half-width are obtained as

$$\sigma_{\mathrm{agg},j} = \sqrt{v_{\mathrm{agg},j}}, \qquad h^{95}_{\mathrm{agg},j} = 1.96\,\sigma_{\mathrm{agg},j}. \tag{11}$$

These quantities summarize the uncertainty that is introduced solely by averaging a correlated meter-level process over a finite segment.

3.5. Slow-sampling stress design

To examine how aggregation uncertainty changes with the sampling design, a controlled slow-sampling stress experiment is performed. The physical segment definitions are kept fixed, while the effective meter-level CT sampling density is reduced by downsampling. The stress levels used in the present study correspond to downsampling steps 1, 2, and 4.

The downsampling is performed coil-wise using a single sampling phase per coil. For a given downsampling step, each coil is assigned one integer offset in the range $\{0, \ldots, \mathrm{step} - 1\}$. This offset is generated deterministically from the coil identifier and a fixed random seed, so that the same coil always receives the same starting phase for a given experimental run, independently of the order in which the data are processed. The retained meter-level samples are then obtained by starting from that coil-specific offset and keeping every step-th observation thereafter. In this way, the stressed series mimics a slower sensor with a fixed acquisition phase along each coil, while avoiding the systematic bias that would arise if all coils always started from the same meter offset.

After downsampling, the segment targets are recomputed from the retained meter-level CT values within each synchronized segment window. At the same time, the effective sample count n_{eff} is recomputed for every segment as the actual number of downsampled meter-level observations that fall inside that window. Thus, n_{eff} represents the number of measurements that are truly available to form the stressed segment mean under the slower sampling regime. If a segment window contains no retained sample after downsampling, a single observed meter-level value nearest to the lower window boundary is used as fallback, and the corresponding segment is assigned $n_{\mathrm{eff}} = 1$. This ensures that the stressed target remains defined while preserving the interpretation that extremely sparse sampling reduces the information content of the segment to a single observed point.

The dependence structure is then re-estimated on the downsampled training coils, converted into a new lookup for $\widehat{v}_{\mathrm{agg}}(n)$, and propagated to the held-out segments using their stressed n_{eff} values. In this way, the stress experiment isolates the effect of slower sampling on label-side aggregation uncertainty without changing the physical segmentation of the target or the downstream model class.

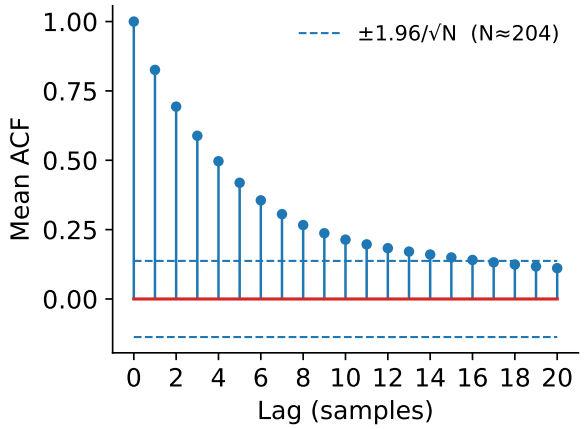
4. RESULTS

Regarding the uncertainty introduced by the aggregation of the CT measurement from the meter level to synchronized segments, the first step is to justify the need for the use of HAC statistics instead of naive i.i.d. scaling. This is presented via the visualization of the ACF metric. To demonstrate how downsampling does not remove the serial dependence of the data, the ACF metric is presented for the downsampled series as well. The impact of accounting for serial dependence on the variance scaling is also presented. Then, to showcase the interpretability and actionability aspect of quantifying the aggregation uncertainty, the half-width of the 95% CI is plotted against the slow-sampled series. For this step, $\gamma(0)$ and n_{eff} are calculated in the training set for both the normal and the downsampled series, and the half-width is calculated on the test set. Finally, the connection between the aggregation uncertainty and the achievable error on a predictor trained on this dataset and deployed on the same test set. Full details of the utilized dataset and the implementation of the predictor are presented in (Kontogiannis et al., 2026).

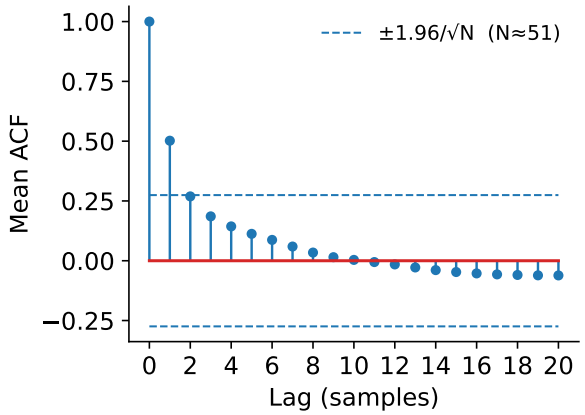
4.1. Serial dependence and variance scaling of aggregated targets

Figure 1 reports the mean autocorrelation function (ACF) of the demeaned meter-level CT series, computed within each coil and averaged across the training coils for both regimes. For visual reference, approximate 95% zero-correlation bounds are shown as $\pm 1.96/\sqrt{N}$, where N denotes the number of observations used for the displayed ACF estimate. In the series with the nominal sampling, the ACF remains strongly positive and decays slowly with lag. Autocorrelation values stay above

the approximate 95% zero-correlation bounds for many lags, indicating persistent short-range dependence in the meter-level fluctuations. In the downsampled series, the ACF is reduced in magnitude and decays more rapidly, approaching zero and becoming slightly negative at higher lags. Nevertheless, the first few lags still exhibit non-negligible positive autocorrelation relative to the 95% reference bounds, indicating that, even for the downsampled series, serial dependence must be accounted for when calculating the variance associated with the aggregation uncertainty.



(a)



(b)

Figure 1. Empirical autocorrelation function (ACF) of the target variable (CT) for a) the base regime and b) the stress regime (downsampling step = 4).

The practical consequence of this dependence is quantified in Figure 2, which shows the variance of the segment mean as a function of the actual segment sample count n, that is, the number of meter-level CT measurements contributing to a synchronized segment target. Under the naive i.i.d. assumption, the aggregation variance decreases rapidly as σ^2/n. In contrast, the dependence-aware estimate obtained from the pooled within-coil HAC statistics decreases more slowly and remains consistently higher across the full range of segment lengths. The shaded band marks the central 5th–95th percentile range of observed segment sample counts in the dataset. Importantly, the discrepancy between the two curves remains substantial throughout this typical operating range, indicating that the uncertainty of most segment-level targets would be underestimated if naive i.i.d. scaling were used. Thus, Figure 2 provides the quantitative bridge between the serial dependence observed in Figure 1 and the uncertainty inflation associated with target aggregation.

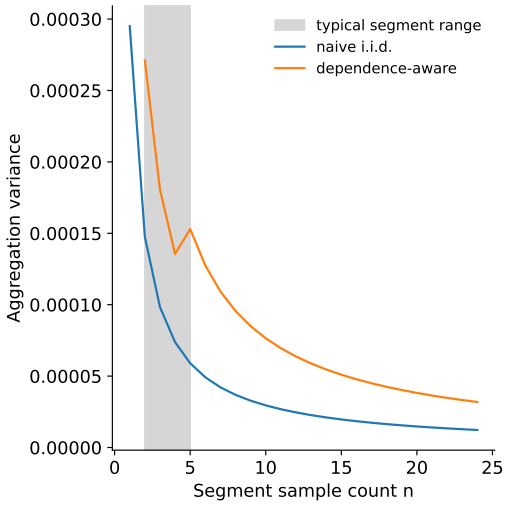


Figure 2. Variance of the segment mean versus segment sample count n under i.i.d. scaling and dependence-aware aggregation variance estimation. The shaded region denotes the central 5th-95th percentile range of observed segment sample counts.

Beyond this baseline characterization, the dependence-aware estimate provides an interpretable handle on how information fidelity changes when the sampling design or the effective process dynamics change. The sampling stress experiment is therefore used next to demonstrate how aggregation uncertainty responds to slower sampling, and how the observed increase in uncertainty can be viewed in reverse as a design lever to increase the achievable predictive performance.

4.2. Slower sampling effect on aggregation uncertainty

Figure 3 reports how aggregation uncertainty changes when the same physical segment windows are observed under progressively slower sampling regimes. For each downsampling step, the meter-level CT series is resampled coil-wise using a fixed phase per coil, the segment targets are recomputed from the retained samples, and the effective sample count n_{eff} is updated accordingly. The plotted quantity is the mean 95% confidence-interval (CI) half-width implied by the dependence-aware aggregation variance on the held-out segments.

A clear monotonic increase is observed as sampling becomes slower. The mean 95% CI half-width increases from approxi-

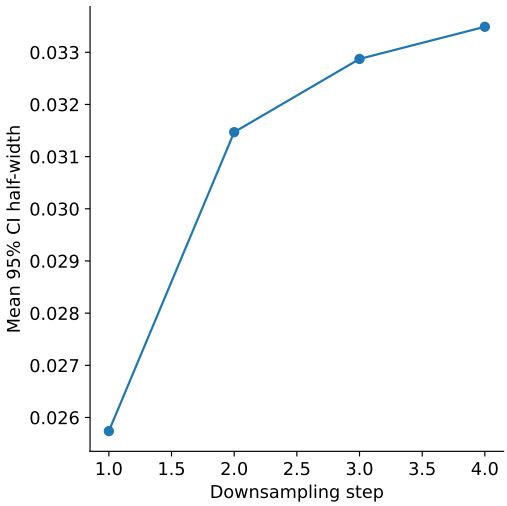


Figure 3. Aggregation uncertainty summary versus sampling design (for example, downsampling step) at fixed physical aggregation window length.

mately 0.0258 at the nominal sampling rate to approximately 0.033–0.034 at step $= 4$. The largest change occurs between step $= 1$ and step $= 2$, after which the increase continues more gradually. This behavior is consistent with the stress design: once fewer meter-level observations are retained within the same synchronized segment, the stressed segment mean is formed from a smaller effective sample count, while the remaining observations still retain short-range dependence. As a result, the information content per segment is reduced and the aggregation-induced uncertainty increases.

The importance of this result is twofold. First, it shows that the uncertainty attached to the target responds directly and interpretabliy to changes in the sampling regime, rather than being a purely abstract statistical quantity. Second, it highlights sampling frequency as a practical design lever. If tighter targets or improved predictive performance are required downstream, increasing the effective sampling density is expected to reduce label-side uncertainty before any changes to the predictive model are considered. This connection between sampling design and target uncertainty is used next to relate aggregation uncertainty to the achievable error scale of segment-level predictors.

4.3. Aggregation uncertainty and the achievable error scale of segment-level predictors

Figure 4 relates the label-side aggregation uncertainty to the error scale achieved by a segment-level CT predictor. The dashed horizontal line shows the test RMSE of 0.0137, achieved with a GBDT and a residual neural network correction after extensive optimization of both architecture and hyperparameters as presented in (Kontogiannis et al., 2026). In additional experiments with alternative model classes, including recurrent and attention-based architectures, the RMSE remained at a comparable level, suggesting that further reductions are difficult to achieve with the dataset at hand. The reason behind this is revealed with the boxplot in Figure 4, which summarizes the dependence-aware aggregation standard deviation σ_{agg} computed on the test set, with a median of $\sigma_{\mathrm{agg}} = 0.0134$. σ_{agg} characterizes the variability introduced upstream when meter-level measurements are converted into segment-level targets under serial dependence.

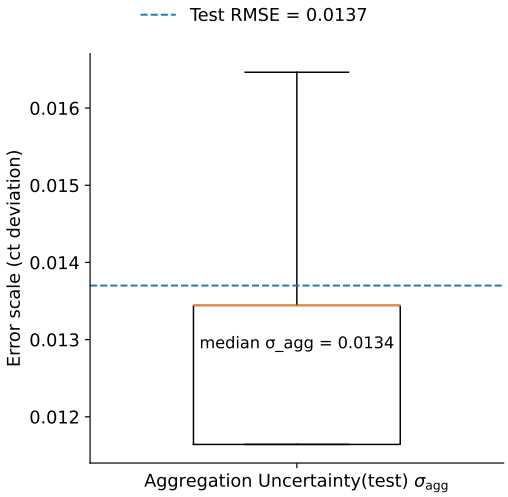
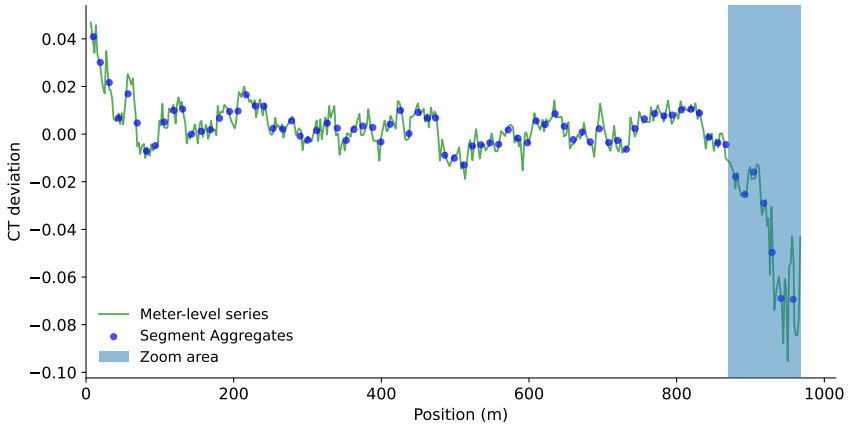


Figure 4. Comparison of predictor RMSE with the aggregation uncertainty scale on the evaluation set.

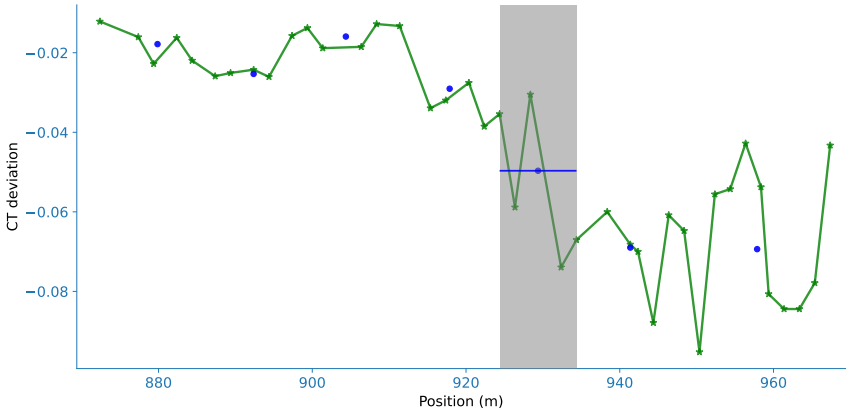
Figure 5 provides an intuitive illustration of this phenomenon by contrasting the meter-level coiling temperature series of an example coil with the segment-level aggregates used as prediction targets. In the overview (Figure 5a), while the aggregated targets follow the local level of the meter-level series, naturally, they suppress within-segment fluctuations. The zoomed view (Figure 5b) emphasizes that within a single segment window (shaded), the meter-level series exhibits substantial variation, whereas the data used as the training target is provided only through one aggregated value (horizontal line). This observation is consistent with the comparison in Figure 4, where the achieved test RMSE is on the same order as the typical σ_{agg}. Thus, when a predictor is trained on these targets, the mapping from features to the observed segment mean is learned through noisy realizations of the same underlying local process. Consequently, even for identical or highly similar feature vectors, the observed aggregated target can plausibly vary on the order of σ_{agg}, and predictions within this scale are deemed as correct by the predictor.

5. CONCLUSIONS

In this study, the value of quantifying the uncertainty introduced when aggregated series are used as targets for a predictive task is discussed. In industrial datasets, the quantity



(a)



(b)

Figure 5. Example coil: meter-level CT deviation series (green) and segment-level aggregated targets (blue). (a) Overview along the strip. (b) Zoomed view highlighting meter-level fluctuations within a segment window (shaded) and the corresponding aggregated segment value (horizontal line).

of interest is often aggregated into coarse segments due to resource constraints or synchronization requirements. While this may appear to be an innocent preprocessing or data-handling step, it has important implications for downstream predictive modelling by directly affecting the achievable performance.

The following conclusions can be drawn:

- Aggregating noisy series into coarse target segments introduces uncertainty directly at the label level, and this uncertainty affects the performance that can realistically be achieved by downstream predictive models.
- Quantifying the uncertainty induced by aggregation is non-trivial, since serial dependencies in the underlying series, as well as real-world deployment constraints, must be taken into account.
- Treating such datasets as deterministic hinders stakeholders from properly understanding their limitations and from setting achievable expectations for predictive performance.

Potential extensions of this work include the consideration of additional label-side uncertainty sources that are likely to be present in industrial datasets with downsampled targets, such as boundary jitter in the considered segmentation of the target value. Their quantification could provide further insight into the total uncertainty of the underlying process and its effect on associated predictive tasks.

REFERENCES

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *59*(3), 817–858. doi: 10.2307/2938229

Deng, J., Sun, J., Peng, W., Hu, Y., & Zhang, D. (2019). Application of neural networks for predicting hot-rolled strip crown. *Applied Soft Computing*, *78*, 119–131. doi: 10.1016/j.asoc.2019.02.030

Fontana, M., Zeni, G., & Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges.

Bernoulli, *29*(1), 1–23. doi: 10.3150/21-BEJ1447

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., … Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(1), 1513–1580. doi: 10.1007/s10462-023-10562-9

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. doi: 10.1007/s10994-021-05946-3

Kontogiannis, T., Melfo, W., Eleftheroglou, N., & Zarouchas, D. (2024). Timeseries feature extraction for dataset creation in prognostic health management: a case study in steel manufacturing. In *Phm society european conference* (Vol. 8, pp. 13–13). doi: 10.36001/phme.2024.v8i1.3968

Kontogiannis, T., Zarouchas, D., & Eleftheroglou, N. (2026). A group-aware temporal framework for quality indicator prediction and anomaly detection in production. *Results in Engineering*, *30*, 110483. doi: 10.1016/j.rineng.2026.110483

Min, K., Kim, K., Kim, S. K., & Lee, D.-J. (2012). Effects of oxide layers on surface defects during hot rolling processes. *Metals and Materials International*, *18*, 341–348.

Mohanty, I., Banerjee, R., Santara, A., Kundu, S., & Mitra, P. (2021). Prediction of properties over the length of the coil during thermo-mechanical processing using DNN. *Ironmaking & Steelmaking*, *48*(8), 953–961. doi: 10.1080/03019233.2020.1848303

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708. doi: 10.2307/1913610

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, *61*(4), 631–653. doi: 10.2307/2297912

Timm, W., Weinzierl, K., Leipertz, A., Zieger, H., & Zouhar, G. (2002). Modelling of heat transfer in hot strip mill runout table cooling. *Steel Research*, *73*, 97-104. doi: 10.1002/SRIN.200200180

Wu, H., Sun, J., Peng, W., & Zhang, D. (2023). Analytical model for temperature prediction of hot-rolled strip based on symplectic space hamiltonian system. *International Journal of Heat and Mass Transfer*, *213*, 124350. doi: 10.1016/j.ijheatmasstransfer.2023.124350

Xie, H. B., Jiang, Z. Y., Liu, X. H., Wang, G. D., & Tieu, A. K. (2006). Prediction of coiling temperature on run-out table of hot strip mill using data mining. *Journal of Materials Processing Technology*, *177*(1-3), 121–125. doi: 10.1016/j.jmatprotec.2006.04.089

Xie, Q., Suvarna, M., Li, J., Zhu, X., Cai, J., & Wang, X. (2021). Online prediction of mechanical properties of hot rolled steel plate using machine learning. *Materials & Design*, *197*, 109201. doi: 10.1016/j.matdes.2020.109201

Yuasa, Y., Yamane, T., Saito, M., Yoshino, M., Miyai, Y., & Shimizu, R. (1990). New temperature control system for hot strip mill run out table. *IFAC Proceedings Volumes*, *23*(8), 143–148.

Zheng, Y., Li, N., & Li, S. (2013, 1). Hot-rolled strip laminar cooling process plant-wide temperature monitoring and control. *Control Engineering Practice*, *21*, 23-30. doi: 10.1016/J.CONENGPRAC.2012.09.004