

PROBABILISTIC FORECASTING OF ELECTRIC VEHICLE CHARGING IN A TIME-VARYING HIERARCHICAL INFRASTRUCTURE

BY GUILLAUME PRINCIPATO^{1,3,a}, YVENN AMARA-OUALI^{1,3,c}, YANNIG GOUDE^{1,3,d},
BACHIR HAMROUCHE^{3,e} AND JEAN-MICHEL POGGI^{1,2,b}

¹*Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, CNRS, INRIA ,*

^a*guillaume.principato@universite-paris-saclay.fr*

²*Université Paris-Cité , ^bjean-michel.poggi@u-pariscite.fr*

³*Electricité de France R&D , ^cyvenn.amara-ouali@edf.fr; ^dyannig.goude@edf.fr; ^ebachir.hamrouche@edf.fr*

Uncertainty quantification for electric vehicle (EV) charging demand is essential to turn the increasing pressure of EV adoption on power systems into a source of flexibility. We study a large-scale dataset of public charging stations in Scotland, where demand exhibits a hierarchical structure, organized across multiple aggregation levels, from individual charging points to regional and national scales. The charging infrastructure is continuously evolving, with new stations deployed and existing ones occasionally removed, which calls for forecasting and uncertainty quantification methods tailored to this evolving structure. To address these challenges, we combine hierarchical forecasting with recent advances in conformal prediction with missing data. The conformal procedures considered produce valid prediction intervals at every node of the hierarchy and the algorithm CP-MNR-Nested* successfully accounts for the time-varying hierarchical structure of the data.

1. Introduction. The introduction begins by motivating the need for uncertainty quantification in electric vehicle (EV) charging demand forecasting and then turns to the specific challenges posed by EV charging infrastructures, with data collected from ChargePlace Scotland, the EV charging public network in Scotland, serving as a study case. It concludes with a brief overview of related work and the organization of the paper.

The large-scale deployment of electric vehicles is reshaping charging infrastructures and electricity consumption patterns, both in magnitude and in temporal structure, and is expected to accelerate further in the coming years (IEA, 2022). This directly concerns distribution system operators and charging network operators, who must plan for and operate a load that keeps growing in scale, evolves as new charging points are continuously deployed, and must be tracked across multiple levels of aggregation, from individual charge points to entire networks (Muratori, 2018; Powell et al., 2022).

Beyond these challenges, it also creates new opportunities to leverage EV charging as a flexible resource for power system operations. In particular, smart charging has been identified as a key mechanism to help balance supply and demand in the presence of intermittent renewable energy sources (Štogl et al., 2024; Bogdanov and Breyer, 2024). Realizing this potential, however, requires reliable forecasts of EV charging demand (Alaraj et al., 2025; Wang et al., 2025), which are subsequently used as inputs to optimization and control algorithms (Fescioglu-Unver and Aktaş, 2023; Diaz-Londono et al., 2024).

In this context, point forecasts alone are not sufficient, as they only provide a partial description of the energy demand, and fail to capture its variability. Decision-making under uncertainty is ubiquitous in power system operations (Gneiting and Katzfuss, 2014), and EV charging demand is no exception: optimization procedures must account for the best

Keywords and phrases. Electric vehicles, conformal prediction, forecast reconciliation, machine learning.

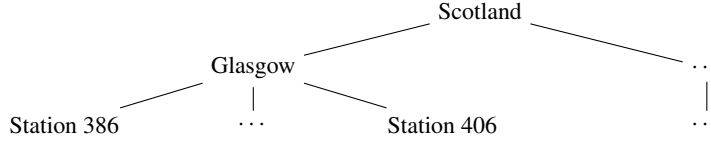


Fig 1: Hierarchical structure of EV charging energy demand in Scotland.

and worst case scenarios to avoid risky decisions that could lead to grid instability or inefficient use of resources (Nutmaki et al., 2024). Probabilistic forecasting has therefore become a standard tool for EV charging demand forecasting (Khan et al., 2023; Zheng et al., 2023; Ostermann and Haug, 2024; Amara-Ouali et al., 2025), with the goal of producing prediction intervals around point forecasts that cover the true values with a specified probability.

EV charging demand forecasting is inherently a multi-resolution problem, as operational decisions are made across a range of spatial and organizational levels, each tied to a distinct market or grid function. For instance, charge point operators need station-level forecasts to schedule maintenance and manage local queuing and on-site pricing; retailers and balance responsible parties need regional aggregates to size day-ahead and intraday market positions; and transmission system operators need national aggregates to dimension reserve requirements and maintain system balance. This range, from individual charging stations to regional aggregates and national systems, has been studied at various granularities (Kim and Kim, 2021; Yi et al., 2022; Ostermann and Haug, 2024).

The forecasting task across these granularities is structurally interdependent, as the energy demand at a given level of aggregation is the sum of the energy demand at the lower levels. This construction induces a hierarchical structure, which can be exploited to share information across the different levels of aggregation to improve the accuracy of forecasts at each level of the hierarchy. For instance, Khan et al. (2023) compare forecasting accuracy across different spatial aggregation levels, from individual charging piles to their sum, and Zhang et al. (2023) study a similar hierarchy for the city of Beijing using large-scale real-world driving and charging data, both for point forecasting; van Etten, Degeler and Luo (2024) propose a global model for point forecasting across multiple stations in four cities, while Zheng et al. (2023) directly exploit the hierarchical structure by enforcing consistency across a network of charging stations in the case of probabilistic forecasting.

In this article, we explore a hierarchical structure within the public charging network in Scotland through data from ChargePlace Scotland (see Figure 1). The data, first studied by Davies, Hart and Galloway (2024), exhibit a rich administrative hierarchical structure, with 1398 charging stations distributed across 32 local authorities, and a national aggregate at the country level. It embodies a central modeling challenge in EV demand forecasting: charging infrastructure evolves over time. The adoption of electric vehicles drives the deployment of new charging stations, while existing stations may be temporarily or permanently removed from the network, e.g., due to maintenance or changes in ownership. Since only active stations contribute to aggregate demand, it induces a so-called time-varying hierarchical structure. These features raise two modeling challenges. The first is to forecast demand for new stations, with limited or no historical data. The second is to account for the current state of the infrastructure, both for point forecasting and for uncertainty quantification. The latter being an important source of variability for EV charging demand.

In addition to these modeling considerations, data from ChargePlace Scotland exhibit several challenging features documented by Davies, Hart and Galloway (2024). Despite a relatively high penetration of electric vehicles in Scotland, total energy demand at the national level shows a decreasing trend over the period of interest (see the right hand side of Figure 2).

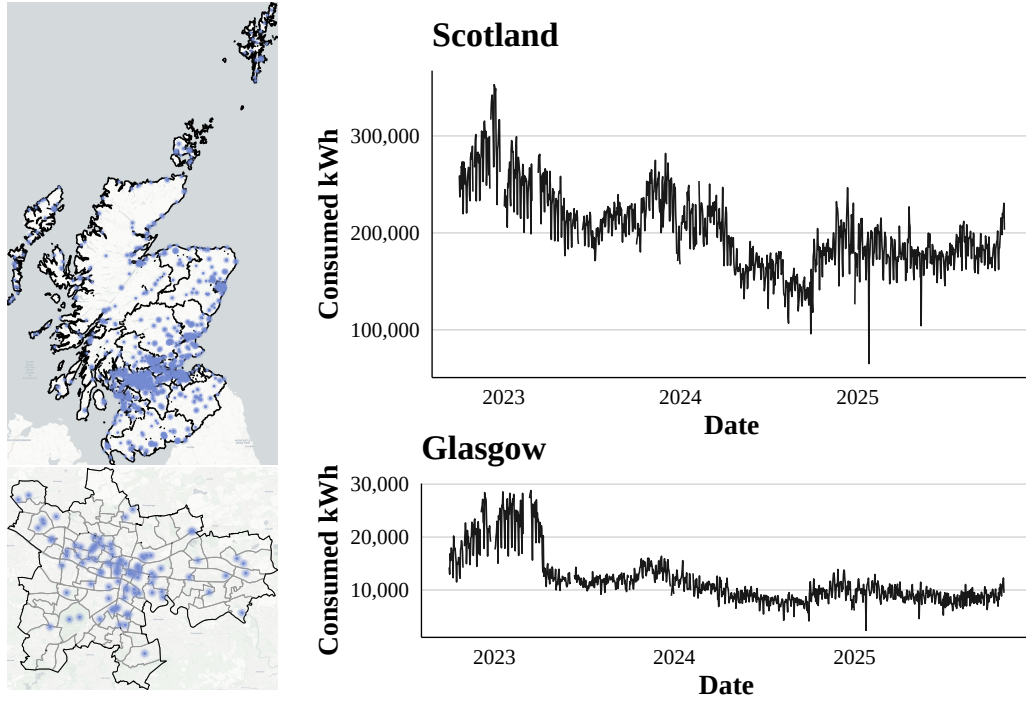


Fig 2: The left panel represents the spatial repartition of the charging stations in Scotland (top) and in Glasgow (bottom). Stations are represented by blue dots, and the size of the dot is proportional to the number of charging points within each station. The right panel shows the corresponding time series of daily energy demand.

This pattern reflects changes in user behavior driven by external factors, including progressive charging tariffs introduced as part of policy incentives. It is particularly visible in the city of Glasgow, where the energy demand dropped on April 4th, 2023, following the introduction of the new pricing scheme in the public charging network (see the right hand side of Figure 2). The introduction of the tariffs leads to a heterogeneous response across the country, as it depends on the local authority, and even on the ownership of the charging station (Davies, Hart and Galloway, 2024). These effects add to other sources of non-stationarity inherent to EV charging demand time series, at both aggregated and disaggregated levels.

Operational forecasting of EV charging demand requires models that not only achieve high predictive accuracy but also satisfy practical constraints such as fast and scalable inference. For tabular data and multivariate time series, it is well known that tree ensembles are highly competitive and easy-to-calibrate forecasting models. For instance, frameworks such as Random Forests (Breiman, 2001), and XGBoost (Chen et al., 2015), are widely used in machine learning competitions (Makridakis, Petropoulos and Spiliotis, 2022; Bojer and Meldgaard, 2021) and industrial applications (Goude, Nedellec and Kong, 2013). Recently, these models have been challenged by deep learning-based approaches. In the context of EV charging, van Etten, Degeler and Luo (2024) consider a global model to forecast energy demand across 55 charging stations in Boulder, Palo Alto, London, and Perth, demonstrating the effectiveness of deep learning architectures like Transformers and N-HiTS (Challu et al., 2023). Alongside deep learning, pre-trained foundation models have driven significant progress in predictive modeling for both tabular and time series data. Models such as TabPFN (Hollmann et al., 2025) and TabICL (Qu et al., 2025) achieve state-of-the-art performance on benchmark tasks like FevBench (Shchur et al., 2025) and TabArena (Erickson et al., 2026). Similar to gradient

boosting machines, these foundation models can be readily adapted to time series forecasting by incorporating time-dependent features or lagged variables. In this work, we utilize TabICL due to its open-source availability and high inference efficiency. However, while these automated machine learning frameworks are highly efficient, many applied settings require domain experts to design interpretable models by integrating physical or systemic knowledge directly into the model equations. This ensures a necessary trade-off between predictive performance and interpretability. Generalized Additive Models (Wood, 2017, GAMs) naturally facilitate this integration, making them highly popular in power system management (Pierrot and Goude, 2011; de Vilmarest et al., 2023; Antoniadis et al., 2024). To achieve a complete and up-to-date overview, we propose considering these three types of forecasting models, along with a simple combination of them, at all three levels of the hierarchy (see Section 5.2).

The aforementioned forecasting methods are tailored for the large number of charging stations in the dataset, but they do not account for the hierarchical structure of the data. This is a common limitation in the context of hierarchical forecasting, which is commonly addressed by producing forecasts that are agnostic to the hierarchical structure, and to subsequently adjust them to ensure that they follow the same aggregation constraints of the data. A well-established framework for this adjustment is forecast reconciliation, introduced by Athanasopoulos, Ahmed and Hyndman (2009); Hyndman et al. (2011). The idea is to combine forecasts from different levels of aggregation through linear combinations to produce coherent forecasts with respect to the hierarchical structure. A natural way to achieve this is through a linear projection of the base forecasts onto a so-called coherent subspace (Paniotelis et al., 2021; Wickramasuriya, Athanasopoulos and Hyndman, 2019). Techniques of forecast reconciliation have been applied in the context of EV charging demand forecasting by Zheng et al. (2023) to produce probabilistic forecasts based on reconciled scenarios.

Recently, conformal prediction has emerged as a particularly attractive framework to quantify uncertainty around point forecasts, as it produces prediction sets with finite-sample and distribution-free coverage guarantees under mild assumptions (Vovk, Gammerman and Shafer, 2005; Lei et al., 2018; Barber et al., 2021). Its appeal in the context of EV charging demand forecasting (Amara-Ouali et al., 2025) lies in its flexibility: similar to forecast reconciliation, conformal prediction can be applied post-hoc to any base predictor, including those obtained from complex machine learning models, without the need for retraining (Angelopoulos and Bates, 2023). This conceptual similarity suggests combining them to produce hierarchy-aware prediction intervals, as proposed by Principato et al. (2026). In particular, the authors show that forecast reconciliation preserves the validity of conformal prediction intervals, while potentially yielding tighter prediction sets, thereby reducing uncertainty around the forecasts. A recent line of work consists in extending conformal prediction to accommodate non-stationarity in time series settings. Gibbs and Candès (2021) introduce adaptive conformal inference, that updates the conformal sets to account for temporal distribution shifts. Adaptive conformal inference has been applied in a variety of energy forecasting settings, including electricity price forecasting by Zaffran et al. (2022), and EV charging demand forecasting by Amara-Ouali et al. (2025).

The rest of the paper is organized as follows. Section 2 provides further details about the dataset and formalizes the forecasting problem. The modeling emphasizes the hierarchical structure, its temporal evolution, and the specific objectives it induces. Section 3 is devoted to a pedagogical exposition of the main ingredients of the methodology, namely forecast reconciliation and conformal prediction in the context of hierarchical forecasting. Section 4 presents the proposed methodology. It describes how to combine forecast reconciliation and conformal prediction in order to account for the time-varying hierarchical structure when quantifying uncertainty. Finally, Section 5 describes the experimental setup and summarizes the performance of the different methods on the ChargePlace Scotland data.

2. EV charging energy demand in Scotland. Data is collected from ChargePlace Scotland, a publicly available record of charging sessions from the largest public EV charging network in Scotland. The resulting dataset is referred to as the ChargePlace Scotland dataset. The code used for the experiments is available on the following [GitHub repository](#).

2.1. The ChargePlace Scotland dataset. The period considered goes from October 2022 to October 2025, extracting session-level information and merging it with spatial and technical characteristics of the charging infrastructure, following [Batic \(2025\)](#). The spatial attributes include the geographical coordinates of each charging point, the address of the charging station, and the local authority to which it belongs. Sessions for which these spatial attributes are missing are excluded from the analysis. Energy consumption at the session level is then aggregated to obtain daily energy demand at each location with available geographic information, thereby constructing a panel of daily demand series at the charging-station level.

The charging stations are indexed by $i = 1, \dots, n$, where $n = 1398$ is the total number of stations. For each day t , let $y_i^{(t)} \in \mathbb{R}$ be the daily energy demand at the i -th station. We consider the natural geographical/administrative partition of the n stations into two levels of aggregation, namely the local authority and the country levels. For instance, the local authority of Glasgow City contains 96 charging stations, and its daily energy demand is thus the sum of the daily energy demand of these 96 stations. This forms a hierarchical structure, which yields a total of $m = 1431$ nodes (1398 stations, 32 local authorities, and 1 national node). The hierarchical structure is illustrated in Figure 1, with the example of the local authority of Glasgow City. This grouping is motivated by the fact that local authorities are the level at which charging infrastructure is planned, funded, and reported by Scottish public bodies (e.g., Transport Scotland, ChargePlace Scotland), making them a natural intermediate level between individual stations and the national system for both operational and policy purposes. This level also coincides with existing administrative and demographic statistics (population, population density, car ownership), which can be used as exogenous covariates.

Nevertheless, the proposed approach can be adapted to other contexts (e.g., by considering finer territorial partitions such as districts or quarters). Let $\mathbf{y}^{(t)} \in \mathbb{R}^m$ be the vector of daily energy demands over all nodes at day t , and consider a multivariate regression setting, where the goal is to forecast demand at every level of the hierarchy from a vector of features $\mathbf{x}^{(t)} \in \mathcal{X}$. This can be formalized as a supervised learning problem, where the objective is to learn a function $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}^m$ that is an estimator of the conditional expectation $\mathbb{E}[\mathbf{y}^{(t)} | \mathbf{x}^{(t)}]$. Vectors of features include weather information, calendar effects, lagged demand, and characteristics of the node of interest (e.g., number of charging points, type of chargers and tariffs indicators), that are known at the time of forecasting.

To model the hierarchical structure of EV charging demand in Scotland, we adopt a standard notation in hierarchical forecasting and encode the aggregation constraints by a binary structural matrix $H \in \{0, 1\}^{m \times n}$. Its entries satisfy $H_{i,j} = 1$ if the j -th station belongs to the i -th node, and $H_{i,j} = 0$ otherwise—see the toy-example depicted in (3). Therefore, for all days t , the vector energy demand at day t belongs to the so-called coherent subspace $\text{Im}(H) = \{H\mathbf{u} : \mathbf{u} \in \mathbb{R}^n\} \subset \mathbb{R}^m$. Forecast reconciliation exploits this property by projecting unconstrained base forecasts, obtained from arbitrary regression models, onto $\text{Im}(H)$ to enforce coherence ([Athanasopoulos, Ahmed and Hyndman, 2009](#); [Hyndman et al., 2011](#)).

2.2. Modeling the infrastructure activity. To capture the temporal evolution of the network, EV charging demand is modeled through a hierarchical structure with a time-varying set of active nodes. The energy demand at some locations—especially at the station level—may be zero during a certain amount of time due to various foreseeable reasons (see Figure 3). For example, if a station does not exist yet, or is temporarily closed due to malfunction,

TABLE 1
Basic notations.

Indices	Domain	Description
t	Days	time index, on a daily basis
i	$[m]$	node index, $i = 1, \dots, m$: stations, $i = 1, \dots, n$ local authorities aggregate, $i = n + 1, \dots, m - 1$ Scotland aggregate, $i = m$
Variables	Domain	Description
$y_i^{(t)}$	$\mathbb{R}$	energy demand at node i and day t
$a_i^{(t)}$	$\{0, 1\}$	infrastructure activity at node i and day t
$\mathbf{x}^{(t)}$	$\mathcal{X}$	vector of features at day t
$\mathbf{y}^{(t)}$	$\mathbb{R}^m$	vector of energy demand at day t
$\mathbf{a}^{(t)}$	$\{0, 1\}^m$	vector of infrastructure activity at day t
H	$\{0, 1\}^{m \times n}$	hierarchical structure of the complete data
$H^{(t)}$	$\{0, 1\}^{m \times n}$	refined hierarchical structure at day t , i.e., with respect to $\mathbf{a}^{(t)}$

renovation work, act of vandalism, or is definitely closed due to change of ownership or de-commissioning. These foreseeable events are conceptually distinct from other zeros that may occur due to the absence of demand, which are essentially concentrated in stations located in rural areas. In the present work, the goal is to leverage the knowledge of the infrastructure activity to improve the accuracy of the forecasts, rather than to forecast when a station will be active or not. Therefore, is it reasonable to assume that the information of whether a node is active (i.e., has non-zero demand) is available at the time of forecasting. We thus consider that the infrastructure activity is known at day t , and focus on forecasting the non-zero coordinates of $\mathbf{y}^{(t)}$. This information is represented by binary vectors of infrastructure activity, defined for each day t as

$$(1) \quad \mathbf{a}^{(t)} \in \{0, 1\}^m, \quad \text{such that} \quad \begin{cases} a_i^{(t)} = 0, & \text{if } y_i^{(t)} = 0, \\ a_i^{(t)} = 1, & \text{otherwise.} \end{cases}$$

We distinguish the vector of infrastructure activity from the vector of features $\mathbf{x}^{(t)} \in \mathcal{X}$. In particular, infrastructure activity is used only within the learning procedure to exclude inactive nodes from the training data, since these correspond to structural zeros that are not targets of interest and may otherwise introduce bias in the resulting forecasts. It is also used post-hoc to refine the hierarchical constraints of the data at day t . We now formalize this refinement of the hierarchical structure.

Inactive nodes do not contribute to the aggregate demand, which implies that enforcing coherence with respect to the original hierarchy may introduce artificial constraints. This motivates the definition of a refined structural matrix and the associated refined coherent subspace for each day t . The refinement of the coherent subspace removes the inactive components and defines a lower-dimensional coherent subspace that better reflects the operational structure on a given day. Precisely, the vector of energy demand at day t belong to the potentially smaller coherent subspace $\text{Im}(H^{(t)}) \subseteq \text{Im}(H)$, where $H^{(t)}$ is defined as follows:

$$(2) \quad H_{i,j}^{(t)} = \begin{cases} 0, & \text{if } a_i^{(t)} = 0, \\ 0, & \text{if } \exists i' \in [m], \text{ s.t. } a_{i'}^{(t)} = 0 \text{ and } H_{i',j} = 1, \\ H_{i,j}, & \text{otherwise.} \end{cases}$$

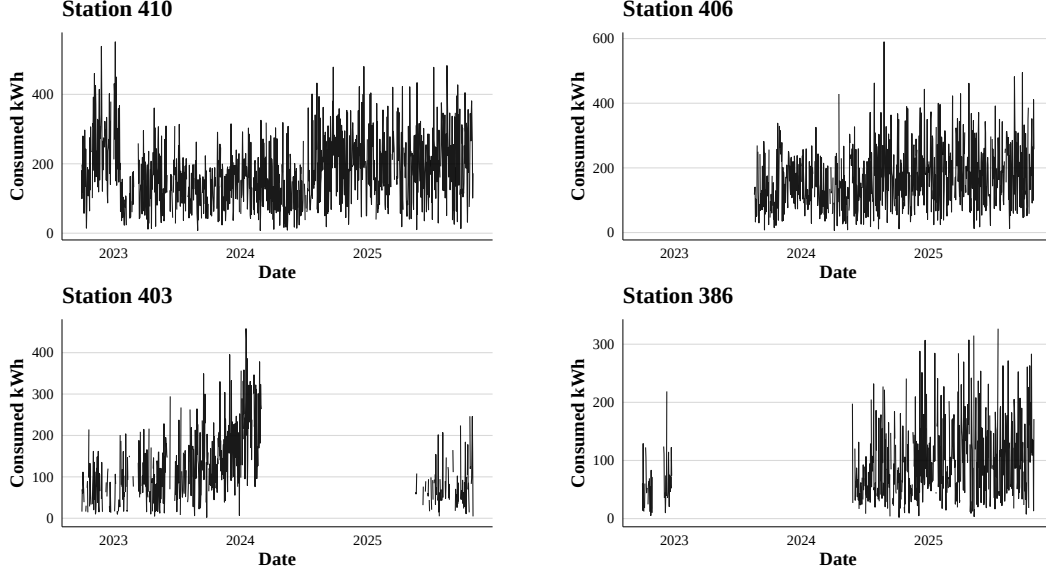


Fig 3: Time series of daily energy demand at four charging stations in Glasgow City. The figure illustrates three typical node behaviors: Station 410 is active throughout the entire period, Station 406 becomes inactive at the end of 2023, Stations 403 and 386 both exhibit the effect of predictable temporary closures.

For example, consider a simplified hierarchy with only three stations and one region. Then, if the vector of infrastructure activity at day t is $\mathbf{a}^{(t)} = (1, 0, 1, 1)^\top$, the following refinement of the structural matrix holds.

$$(3) \quad H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{may be refined into} \quad H^{(t)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

2.3. Coverage objectives. Infrastructure activity is a key driver of EV charging demand and must therefore be accounted for when assessing forecast uncertainty. This is addressed by extending the usual marginal coverage requirement to a conditional setting.

The goal is to produce reliable forecasts of EV charging demand at each node of the hierarchy, i.e., for each station, local authority, and the national aggregate, on a time period indexed by $t \in \text{Test}$, and to quantify the uncertainty around these forecasts. Given a nominal coverage level $1 - \alpha = 90\%$, a classic requirement is to produce prediction intervals at node i containing the true demand for at least 90% of the days in Test. For instance, if infrastructure activity is properly accounted for (e.g., through forecast reconciliation), there is no uncertainty about the prediction of energy demand at an inactive station. Hence, marginal coverage may appear satisfactory simply because inactive periods are trivially covered, while coverage on active periods remains poor. Similarly, the uncertainty quantification at an active station may depend on the activity of other stations in the hierarchy. For example, a temporary closure of neighboring stations may lead to a surge in demand at the station under consideration, resulting in a higher variability of the energy demand. As a consequence, assessing coverage by averaging over all days in Test may not be representative of the underlying uncertainty.

A natural solution is to extend the coverage requirement to a conditional one, where coverage is assessed by averaging over subsets of Test corresponding to each specific configuration

of the infrastructure. This is infeasible in practice, as the number of possible configurations is large and only few observations are available for each. Therefore, we consider two weaker conditional coverage requirements, where coverage is assessed by averaging over subsets of Test corresponding to ranges of similar infrastructure activity patterns. The first is a local conditioning, where coverage is assessed by averaging over the days for which the node under consideration is active. This is the minimal requirement for a meaningful coverage assessment, as we are only interested in forecasting the energy demand at active nodes. In the rest of the article, this condition is referred to as validity. The second is a stronger conditioning, where coverage is assessed by averaging over the days for which the node under consideration is active, and the number of active nodes in its neighborhood—defined by the local authority—partitions the time indices into two subsets: low and high activity regimes. The rationale for this additional conditioning is that the activity of the nodes within the same local authority is an important source of variability for the accuracy of the forecasts. Note that other conditioning could be considered, for instance by changing the definition of the neighborhood, or the number of activity regimes. This second requirement is referred to as conditional coverage or, equivalently, conditional validity.

2.4. Efficiency objectives. Beyond validity, the intervals should also be efficient, in the sense that they should be as narrow as possible. Indeed, intervals that are too wide are valid as they always contain the true value, but will not be representative of the underlying uncertainty and will not be useful for decision-making. For instance, the prediction interval $[0 \text{ kWh}, 10000 \text{ kWh}]$ for the energy demand at a station will always contain the true value, but the station operator should not prepare for a demand of 10000 kWh that is not realistic. Therefore, the efficiency of the prediction intervals is measured by their average length. As for validity, the average length is computed only over the days for which the node under consideration is active.

3. Point and probabilistic forecasting. Forecast reconciliation and conformal prediction are two post-hoc frameworks that can be combined to forecast hierarchical data, such as EV charging demand and provide the associated prediction intervals. We first review the two forecast reconciliation procedures considered in this article, in the context of point forecasting, and then describe how they can be integrated within conformal prediction.

3.1. Two forecast reconciliation procedures. We use OLS reconciliation (ordinary least squares, [Hyndman et al., 2011](#)) and introduce rOLS (refined OLS), its natural extension that accounts for the time-varying hierarchical structure modeling introduced in Section 2.2. Both procedures perform reconciliation through orthogonal projection: OLS projects base forecasts onto the fixed coherent subspace $\text{Im}(H)$, whereas rOLS projects them, at day t , onto the refined coherent subspace $\text{Im}(H^{(t)})$.

Given a vector of base forecasts $\hat{\mathbf{y}}^{(t)} \in \mathbb{R}^m$, the following closed-form expressions hold for the reconciled forecasts obtained from OLS and rOLS, respectively,

$$(4) \quad \tilde{\mathbf{y}}_{\text{OLS}}^{(t)} = H(H^\top H)^{-1}H^\top \hat{\mathbf{y}}^{(t)}, \quad \text{and} \quad \tilde{\mathbf{y}}_{\text{rOLS}}^{(t)} = H^{(t)}(H^{(t)\top}H^{(t)})^\dagger H^{(t)\top} \hat{\mathbf{y}}^{(t)},$$

where $\dagger$ denotes the Moore-Penrose pseudo-inverse.

The impact of using rOLS instead of OLS is illustrated on the toy-example, with $n = 3$ stations, $m = 4$ nodes, the structural matrix given in (3), and the infrastructure activity vector $\mathbf{a}^{(t)} = (1, 0, 1, 1)^\top$ at day t . The following numerical example of EV charging demand and base forecasts, induces the corresponding OLS and rOLS reconciled forecasts:

$$(5) \quad \mathbf{y}^{(t)} = \begin{bmatrix} 26 \\ 0 \\ 32 \\ 58 \end{bmatrix}; \quad \hat{\mathbf{y}}^{(t)} = \begin{bmatrix} 20 \\ 20 \\ 30 \\ 65 \end{bmatrix}; \quad \tilde{\mathbf{y}}_{\text{OLS}}^{(t)} = \begin{bmatrix} 18.75 \\ 18.75 \\ 28.75 \\ 66.25 \end{bmatrix}; \quad \tilde{\mathbf{y}}_{\text{rOLS}}^{(t)} = \begin{bmatrix} 25 \\ 0 \\ 35 \\ 60 \end{bmatrix}.$$

TABLE 2
Advanced notations.

Functions	Domain	Description
$\hat{A}$	$\mathcal{X} \times \{0, 1\}^m \rightarrow \mathbb{R}^m$	base predictor
$\tilde{A}$	$\mathcal{X} \times \{0, 1\}^m \rightarrow \mathbb{R}^m$	reconciled predictor
$\hat{R}$	$\mathbb{R}^m \times \{0, 1\}^m \rightarrow \mathbb{R}^m$	reconciliation function
Variables	Formula	Description
$\tilde{\mathbf{y}}^{(k)} \in \mathbb{R}^m$	$\tilde{A}(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}) = \hat{R}(\hat{A}(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}), \mathbf{a}^{(k)})$	reconciled forecasts at day k
$\mathbf{s}^{(k)} \in \mathbb{R}^m$	$ \tilde{\mathbf{y}}^{(k)} - \mathbf{y}^{(k)} = s((\mathbf{x}^{(k)}, \mathbf{a}^{(k)}), \mathbf{y}^{(k)}; \tilde{A})$	vectorial non-conformity score at day k
$\tilde{y}_i^{(k)} \in \mathbb{R}$	$\Pi_i(\tilde{\mathbf{y}}^{(k)})$	reconciled forecasts for node i at day k
$s_i^{(k)} \in \mathbb{R}$	$ \tilde{y}_i^{(k)} - y_i^{(k)} $	non-conformity score for node i at day k

Apart from the second coordinate, which is zero and thus excluded from the forecasting task, the rOLS reconciled forecasts are closer to the true values than both the base forecasts and the OLS reconciled forecasts. Indeed, OLS propagates non-zero value produced by the base forecast at the inactive station. It affects the other coordinates and, in particular, leads to an inflated last coordinate relative to the true value. By contrast, rOLS constructs a reconciled vector in such a way that none of its components depends on the second coordinate of the base forecast, which is effectively set to zero, and remaining coordinates are adjusted accordingly.

3.2. Basics on conformal prediction for hierarchical data. Conformal prediction for hierarchical data (Principato et al., 2026) is a natural framework to quantify uncertainty in the context of hierarchical forecasting. Originally introduced for fixed hierarchical structures, it can easily be adapted to handle time-dependent reconciliation procedures such as rOLS, as described in Section 3.1. We recall the main ingredients of the algorithm, which we refer to as CP (for Conformal Prediction), and introduce the notations that will be used in the subsequent sections. The procedure is illustrated in Figure 4.

CP is a multivariate extension of inductive conformal prediction, also known as split conformal prediction (Vovk, Gammernan and Shafer, 2005; Lei et al., 2018), where prediction sets are constructed component-wise and the procedure include a reconciliation step. The first step of CP consists in splitting the available data into a training and a calibration set, indexed by Train and Cal, respectively. The training set is used to fit a base predictor $\hat{A} : (\mathbf{x}, \mathbf{a}) \rightarrow \hat{\mathbf{y}}$, and a reconciliation function $\hat{R} : (\hat{\mathbf{y}}, \mathbf{a}) \rightarrow \tilde{\mathbf{y}}$, resulting in a reconciled predictor $\tilde{A}(\mathbf{x}, \mathbf{a}) \rightarrow \tilde{\mathbf{y}}$ defined as the combination of the base predictor and the reconciliation function:

$$(6) \quad \begin{aligned} \tilde{A} : \mathcal{X} \times \{0, 1\}^m &\longrightarrow \mathbb{R}^m, \\ (\mathbf{x}, \mathbf{a}) &\longmapsto \hat{R}(\hat{A}(\mathbf{x}, \mathbf{a}), \mathbf{a}). \end{aligned}$$

The calibration set is then used to compute non-conformity scores, that measure how well the reconciled forecasts fit the calibration observations at each node $i \in [m]$. We consider the standard non-conformity score function that consists of the absolute difference between the reconciled forecasts and the true values, i.e.,

$$(7) \quad \begin{aligned} s_i : (\mathcal{X} \times \{0, 1\}^m) \times \mathbb{R}^m \times \mathcal{F} &\longrightarrow \mathbb{R}, \\ ((\mathbf{x}, \mathbf{a}), \mathbf{y}; \tilde{A}) &\longmapsto |y_i - \tilde{y}_i| = |\Pi_i(\mathbf{y} - \tilde{A}(\mathbf{x}, \mathbf{a}))|, \end{aligned}$$

where $\mathcal{F} = \{f : \mathcal{X} \times \{0, 1\}^m \rightarrow \mathbb{R}^m\}$ is the set of predictors, and Π_i is the operator that extracts the i -th component of a vector in $\mathbb{R}^m$.

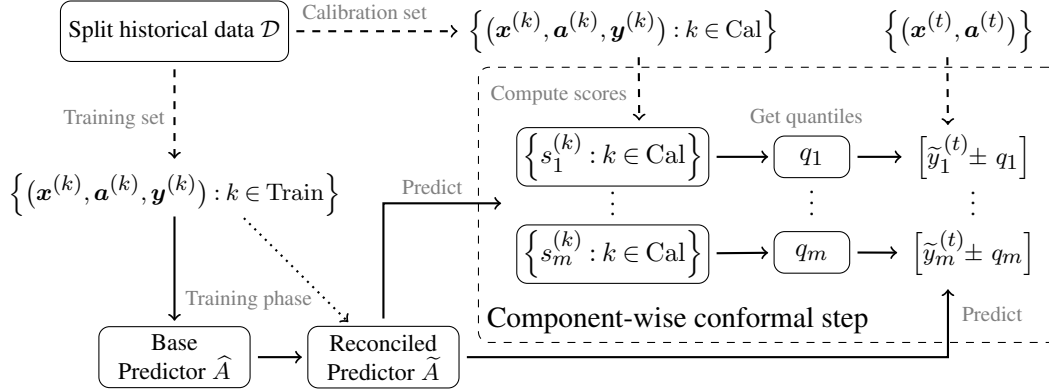


Fig 4: The CP procedure, adapted from conformal prediction for hierarchical data.

Denoting by $s_i^{(k)}$ the non-conformity score for node i at day $k \in \text{Cal}$, the $(1 - \alpha)$ -empirical quantile of the calibration non-conformity scores is computed, up to the usual finite-sample correction, namely,

$$(8) \quad q_i = \min \left\{ q \in \mathbb{R} : \# \{ k \in \text{Cal} : s_i^{(k)} \leq q \} \geq (1 - \alpha)(\# \text{Cal} + 1) \right\},$$

where the cardinality of a set S is denoted by $\#S$.

Finally, at each node $i \in [m]$, the resulting conformal set is the prediction interval centered around $\tilde{y}_i^{(t)}$, the i -th component of the reconciled forecasts for the new observation, and with a radius equal to the empirical quantile q_i , i.e.,

$$(9) \quad \forall i \in [m], \quad \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) = [\tilde{y}_i^{(t)} \pm q_i].$$

4. Conformal prediction for time-varying hierarchical structures. The conformal procedure described in Section 3.2 encompasses forecast reconciliation methods that account for the infrastructure activity, but the construction of the prediction intervals is not adapted to the time-varying nature of the hierarchy. In particular, the method is designed to produce prediction intervals that are—only—valid marginally over all possible configurations of the infrastructure activity for the test day t . This is because the non-conformity scores are computed over the entire calibration set, which includes both active and inactive nodes. It is problematic in the case of EV charging demand, as the non-conformity scores for inactive nodes stations trivially zero for a reconciled predictor based rOLS, leading to prediction intervals that are too small for active nodes that have been inactive during the calibration period.

This issue is addressed in the subsequent sections by proposing three extensions of the CP procedure. The first extension, referred to as CP-MNR (for Conformal Prediction with Missing Nodes Reconciliation), is designed to produce prediction intervals that are valid conditionally on the node being active, a minimal requirement for a meaningful coverage assessment. The second extension, CP-MNR-Nested*, is designed to produce prediction intervals that are valid conditionally on the entire configuration of the infrastructure activity, a more ambitious goal. Finally, the third extension, Adaptive CP-MNR, is designed to produce prediction intervals that are valid conditionally on the node being active, even when the data exhibit a distribution shift between the calibration and test points.

4.1. Prediction intervals for active nodes. CP-MNR is a simple extension of CP, designed to produce prediction intervals that are valid conditionally on the node being active.

Algorithm 1 CP-MNR at level $1 - \alpha$

Input: Dataset $\mathcal{D} = \{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)}) : k \in \text{Train} \cup \text{Cal}\}$; algorithms $\mathcal{A}$ and $\mathcal{R}$; test point $(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$.

1: Fit the predictors on $\{\mathcal{D}; \text{Train}; \mathcal{A}; \mathcal{R}\}$ and get $\{\hat{A}(\cdot, \cdot), \hat{R}(\cdot, \cdot), \tilde{A}(\cdot, \cdot)\}$

for $i \in [m]$ **do** perform Mondrian conformal prediction:

2: Consider a subsampled set of calibration indices $\widetilde{\text{Cal}}(i) = \{k \in \text{Cal} : a_i^{(k)} = a_i^{(t)}\}$

3: Compute the non-conformity scores, for $k \in \widetilde{\text{Cal}}(i)$, $s_i^{(k)} = s_i((\mathbf{x}^{(k)}, \mathbf{a}^{(k)}), \mathbf{y}^{(k)}; \tilde{A}(\cdot, \cdot))$

4: Compute their $(1 - \alpha)$ -empirical quantile, defined as

$$\tilde{q}_i = \min \left\{ q \in \mathbb{R} : \# \left\{ k \in \widetilde{\text{Cal}}(i) : s_i^{(k)} \leq q \right\} \geq (1 - \alpha)(\#\widetilde{\text{Cal}}(i) + 1) \right\}$$

5: Build the prediction interval, $\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) = [\Pi_i(\tilde{A}(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})) \pm \tilde{q}_i]$

end for

Output: Predictions intervals $\{\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) : i \in [m]\}$.

It only differs from the CP procedure in that the non-conformity scores are computed on a subsampled set of calibration indices. For each node $i \in [m]$, the subsampled set of calibration indices is denoted by $\widetilde{\text{Cal}}(i) \subseteq \text{Cal}$ and is defined as the subset of calibration indices corresponding to observations that share the same activity status of the i -th node as the test point, namely,

$$(10) \quad \forall i \in [m], \quad \widetilde{\text{Cal}}(i) = \{k \in \text{Cal} : a_i^{(k)} = a_i^{(t)}\}.$$

The rest of the procedure remains unchanged (i.e., follows the steps described in Section 3.2). For each node $i \in [m]$, non-conformity scores are computed on the restricted calibration set, empirical quantile at level $1 - \alpha$ of these scores is computed, and the resulting prediction interval is defined based on this quantile as in (9).

The CP-MNR procedure described above is detailed in Algorithm 1. Its construction is reminiscent of the Mondrian conformal prediction framework (Vovk, Gammernan and Shafer, 2005), which aims to achieve validity conditional on a categorical variable (here the binary activity of the node under consideration). The theoretical guarantees of CP-MNR are therefore similar to those of Mondrian conformal prediction, and are stated in the Supplementary Material A.2. The fitting of the predictors involved in the first step of Algorithm 1 is stated formally in the Supplementary Material B.1,

REMARK 1. In the case of a node $i \in [m]$ that is always active (e.g., an aggregated node), the CP-MNR procedure coincides with the CP procedure of Section 3.2. In Section 5, CP is thus omitted from the numerical experiments for the aggregated nodes.

4.2. Infrastructure-aware prediction intervals. As discussed in Section 2.3, conditioning on the node being active is a minimal but not sufficient requirement as whole infrastructure activity is also an important source of variability for the accuracy of the forecasts. In this section, the CP-MNR procedure is extended to produce prediction intervals that are valid conditionally on the entire configuration of the infrastructure activity.

In principle, infrastructure activity vector is also a categorical variable so that Mondrian conformal prediction could be applied. Nevertheless, this would require to partition the calibration set into 2^m subsets, one for each possible configuration of the infrastructure activity, which exceeds the number of available observations by many orders of magnitude. This practical limitation advocates for a procedure that accounts for the whole infrastructure activity without requiring a separate calibration set for each configuration: CP-MNR-Nested*.

Before introducing CP-MNR-Nested*, for the sake of a more pedagogical exposition, we first rewrite the conformal step of CP-MNR in a equivalent form that emphasizes the role of infrastructure activity in the construction of the prediction intervals. For all node $i \in [m]$, the i -th prediction interval output by Algorithm 1 can be equivalently written as the set of candidate values whose associated non-conformity score is strictly larger to at most a fraction $(1 - \alpha)$ of the non-conformity scores computed on the calibration set, up to the usual finite-sample correction, i.e.,

$$\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) = \left\{ \Pi_i(\mathbf{y}) \in \mathbb{R} : (1 - \alpha)(\#\widetilde{\text{Cal}}(i) + 1) > \# \left\{ k \in \widetilde{\text{Cal}}(i) : s_i((\mathbf{x}^{(k)}, \mathbf{a}^{(k)}), \mathbf{y}^{(k)}; \tilde{A}) < s_i((\mathbf{x}^{(t)}, \mathbf{a}^{(t)}), \mathbf{y}; \tilde{A}) \right\} \right\}.$$

This formulation highlights that the non-conformity scores computed on the calibration set only depend on the infrastructure activity of the test point through the selection of the calibration indices, which is a local conditioning on the activity of the node under consideration.

The rationale of CP-MNR-Nested* is to enforce a symmetric treatment of calibration and test points with respect to the infrastructure activity, by evaluating non-conformity scores under a shared activity pattern. The resulting procedure is stated in Algorithm 2, and the symmetric treatment of calibration and test points is highlighted in blue. We now describe the three steps of the procedure in more details.

The first step of CP-MNR-Nested* is identical to that of CP-MNR, where the predictors are fitted on the training set. In the second step, for each calibration index $k \in \text{Cal}$, a common augmented infrastructure activity pattern $\tilde{\mathbf{a}}^{(k)} = \min(\mathbf{a}^{(k)}, \mathbf{a}^{(t)})$ is considered, where the minimum is taken component-wise. It corresponds to a hypothetical scenario, where the only active nodes are those that are active in both the calibration and test points.

The third step consists in computing the non-conformity scores of the augmented calibration points $(\mathbf{x}^{(k)}, \tilde{\mathbf{a}}^{(k)}, \mathbf{y}^{(k)})$ to compare them with those of the temporary test points $(\mathbf{x}^{(t)}, \tilde{\mathbf{a}}^{(k)}, \mathbf{y})$, for all candidate values $\mathbf{y} \in \mathbb{R}^m$. This comparison, performed under the same augmented infrastructure activity pattern determines the cardinality of the following sets:

$$(11) \quad \forall \mathbf{y} \in \mathbb{R}^m, \quad \left\{ k \in \text{Cal} : s_i((\mathbf{x}^{(k)}, \tilde{\mathbf{a}}^{(k)}), \mathbf{y}^{(k)}; \tilde{A}) < s_i((\mathbf{x}^{(t)}, \tilde{\mathbf{a}}^{(k)}), \mathbf{y}; \tilde{A}) \right\}.$$

It is then used to define the prediction intervals, as in the above reformulation of CP-MNR.

This construction does not describe how to implement the procedure. The Supplementary Material B.2 is devoted to this purpose: it explains how to derive the conformal set, why a subsampled set of calibration indices $\widetilde{\text{Cal}} \subseteq \text{Cal}$ should be used, and discusses the nature of the resulting prediction set as well as the computational cost of the procedure.

The construction of CP-MNR-Nested* is inspired from the missing data augmentation (MDA) procedure of Zaffran et al. (2023, 2024) and is actually an extension to our setting of their algorithm, CP-MDA-Nested* (Zaffran et al., 2024, Algorithm 1). We establish a connection between the two frameworks in the Supplementary Material A.1, and show that the theoretical guarantees of CP-MDA-Nested* can be extended to CP-MNR-Nested* (see the Supplementary Material A.3). Even though some of the assumptions that are natural for the former are not satisfied in the latter, the theoretical guarantees of CP-MNR-Nested* provide the heuristic behind the construction of the procedure.

4.3. Online extension of the methodology. The procedures described in Sections 3.2, 4.1 and 4.2 (CP, CP-MNR and CP-MNR-Nested*) are designed for a single test point. The underlying assumption behind their construction follows the standard conformal prediction

Algorithm 2 CP-MNR-Nested* at level $1 - \alpha$

Input: Dataset $\mathcal{D} = \{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)}) : k \in \text{Train} \cup \text{Cal}\}$; algorithms $\mathcal{A}$ and $\mathcal{R}$; test point $(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$; subsampled set of calibration indices $\widetilde{\text{Cal}} \subseteq \text{Cal}$.

1: Fit the predictors on $\{\mathcal{D}; \text{Train}; \mathcal{A}; \mathcal{R}\}$ and get $\{\widehat{A}(\cdot, \cdot), \widehat{R}(\cdot, \cdot), \widetilde{A}(\cdot, \cdot)\}$

2: **for** $k \in \widetilde{\text{Cal}}$ **do** consider an augmented infrastructure pattern
 $\widetilde{\mathbf{a}}^{(k)} = \min(\mathbf{a}^{(k)}, \mathbf{a}^{(t)})$

end for

3: **for** $i \in [m]$ **do** build the prediction set $\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$ defined as

$$\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) = \left\{ \Pi_i(\mathbf{y}) \in \mathbb{R} : (1 - \alpha)(\#\widetilde{\text{Cal}} + 1) > \right. \\ \left. \# \left\{ k \in \widetilde{\text{Cal}} : s_i\left((\mathbf{x}^{(k)}, \widetilde{\mathbf{a}}^{(k)}), \mathbf{y}^{(k)}; \widetilde{A}(\cdot, \cdot)\right) < s_i\left((\mathbf{x}^{(t)}, \widetilde{\mathbf{a}}^{(k)}), \mathbf{y}; \widetilde{A}(\cdot, \cdot)\right) \right\} \right\}$$

end for

Output: Predictions sets $\{\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) : i \in [m]\}$.

framework, where the main assumption is that the calibration points and the test point are drawn independently from the same distribution. In the context of EV charging demand forecasting, this setting needs to be extended to a sequential setting, defined as follows.

At each round $t \in \text{Test}$, the practitioner observes the covariates and the infrastructure activity vector $(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$, and output for each node i a prediction set $\mathcal{C}_i^{(t)} = \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$ for the unknown value $y_i^{(t)}$, which is then revealed at the end of the round. In this setting, no stochastic assumption is made on the sequence of test points to account for the temporal dependence and distributional shifts that are inherent to EV charging demand data, as illustrated in Figures 2 and 3.

These attributes advocate for an online extension of the conformal procedures through adaptive conformal inference (ACI, Gibbs and Candès, 2021). The principle of the adaptive extension is to update the conformal intervals sequentially based on the recent miscoverage events. For each node $i \in [m]$, the fixed nominal level $1 - \alpha = 90\%$ used in the conformal step is replaced by a node-specific level, updated sequentially over the test period. If node i is inactive at the current test point, the coverage level is left unchanged. Otherwise, the coverage level increases if the previous conformal interval did not cover the true value, leading to a wider interval, and decreases if the previous conformal interval did cover the true value, leading to a narrower one.

This strategy can be used to extend each of the three conformal procedures described in the previous sections. Nevertheless, the coverage level is updated only for active nodes, which is consistent with the construction of the CP-MNR procedure that is designed to account for the activity of the node under consideration, but not with the construction of CP and of CP-MNR-Nested*. In particular, this adaptive extension may hinder the stronger conditional coverage guarantee targeted by the CP-MNR-Nested* procedure. Therefore, only the adaptive extension of CP-MNR is considered, and is referred to as Adaptive CP-MNR. Its formal statement and the proof of its theoretical guarantees are deferred to the Supplementary Material B.3.

5. Forecasting EV charging energy demand with confidence. This section describes the study case of EV charging demand forecasting on the ChargePlace Scotland data and depict how the proposed methodology addresses the specific challenges of this application. We first introduce the experimental setting, including the rolling-window protocol and the data preprocessing choices. Then, the most appropriate base predictor and reconciliation strategy are selected and fixed for the subsequent experiments. Finally, the performance of the proposed conformal procedures is evaluated and analyzed.

5.1. Experimental setting. The proposed methodology is assessed on a six-month test period, denoted by *Test*, from February 5th to August 18th, 2024. The period exhibits a significant temporal drift in the data distribution (Figure 2), making it a particularly challenging setting for forecasting. As discussed in Section 4.3, the CP, CP-MNR, and CP-MNR-Nested* procedures are designed for a single test point. To deploy them sequentially over the entire test period, one could in principle retrain the underlying models at each time step, at a substantial computational cost, or, at the other extreme, keep the models fixed throughout the whole period. We adopt an intermediate strategy and update the models every two weeks to account for the dynamic nature of the data. Accordingly, the training and calibration periods are updated using a rolling-window scheme, illustrated in Figure 5.

The variability of the performance metrics is evaluated by running the experiments 10 times with different rolling-window partitions of the historical data. The resulting mean estimates and standard errors are reported throughout this section. For each experimental run, the available data before the first test block are split into a training set and a calibration set. This partition is then shifted forward by two weeks at each subsequent test block. The calibration set represents 20% of the historical data and is constructed as follows: it includes the week preceding the test block together with one-week blocks sampled uniformly at random from the remaining time indices. This partitioning affects the prediction intervals through a change in the predictors fitted on the training set and in the calibration points used to compute the non-conformity scores.

Before fitting the predictors, the data is preprocessed to remove sessions with missing geographical information and, across all nodes, dates with anomalous data reporting as identified by Davies, Hart and Galloway (2024). Subsequently, features are engineered for the regression tasks across the three hierarchical levels. Calendar features encode weekly and yearly cycles using day-of-week and day-of-year variables (scaled between 0 and 1). Public holidays are recoded and treated as Sundays. Weather features are collected by requesting data with the `riem` package (Salmon and Elchison, 2025). Daily maximum temperature and daily average humidity are calculated at each node. At the station level, these meteorological features are derived using a distance-weighted average from weather stations. For regional and national levels, the station level features are aggregated using a weighted average based on the number of charging points per station. With the same weighted average scheme, charger locations (longitude and latitude) are considered both at the station and regional levels. Equipment features are also included, such as the proportion of slow, accelerated, and fast chargers at each node as well as a dimensioning factor for the number of charging points and paid charging points, estimated dynamically using the prior day’s number of used chargers and the maximum number of used chargers over the previous 30 days. Charging cost features account for the transition of chargers from free to paid described in Section 1. Historical energy consumption incorporates the prior day’s energy, the energy consumed on the same day of the previous week, and a 1-to-7-day rolling mean. Days with zero demand are set to NA (i.e., treated as missing values) so that predictors are fitted exclusively on active nodes. To prevent the propagation of these NAs into the lags and thus avoid reducing the number of available observations, any NA in a lagged variable is imputed using the last available observation of the same day type. Finally, depending on the model, a shifted logarithmic transformation is applied to the target and the lags at the regional and station levels. This prevents the model from favoring larger nodes while ensuring numerical stability for nodes with near-zero demand.

5.2. Choice of the base predictor and reconciliation procedure. At all three levels of the hierarchy, we consider a random forest (RF), a generalized additive model (GAM), a (pre-trained) foundation model for time series and tabular data (tabICL), and the combination of

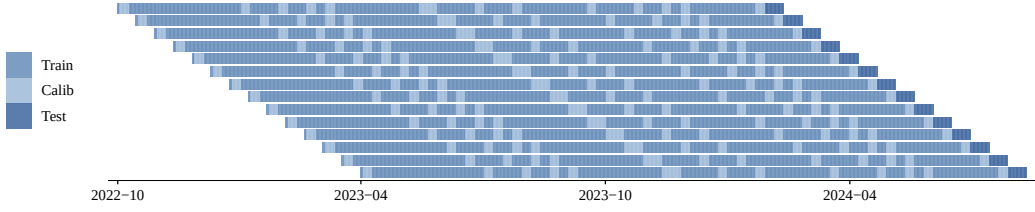


Fig 5: Rolling-window data partitioning. At the first block, the available data is split into a training set (80%) and a calibration set (20%). The calibration set of indices is drawn once, combining the week preceding the test period and randomly sampled one-week historical blocks. The same partition is then shifted forward by two weeks across subsequent blocks.

these models (Combi) that consists of the average of the predictions output by these models. For the regional and station levels, a boosting model (XGB) is added as a base model and included in the average of Combi. For the national level, since there is only a single time series, the models exclusively use calendar, weather, and charging cost features, along with lagged energy. Conversely, at the regional and station levels, due to the large number of nodes, a global modeling approach is considered, where a single model is fitted for all nodes within each aggregation level. In this setting, features that enable cross-learning among series, such as infrastructure information and spatial features, are included. The GAM is highly efficient for load forecasting, as its additive decomposition and non-linear capabilities are well-suited to model the various effects influencing energy consumption (Antoniadis et al., 2024). When applied as a global model, advanced features in GAMs, such as the use of tensor products and combined effects, allow the model to capture the complexity of the data. Furthermore, tree-based models such as RF and XGB are particularly efficient in a global setting, since they naturally group different time series to harness cross-learning. TabICL achieves this same global modeling objective through a different approach, relying on cross-attention mechanisms. An overview of the related work for the different models is provided in Section 1.

REMARK 2. To keep computation time low and ensure our experiments remain reproducible on a standard laptop for the station level, we restrict the tabICL context to 10^5 randomly sampled data points and compute a bagging estimate across 5 TabICL models of the energy demand by batch of 1024 data points. Note that this approach does not degrade predictive performance; rather, it allows us to effectively manage the memory footprint and computation time.

A base predictor is a multivariate predictor that outputs a vector of forecasts for all nodes in the hierarchy, and can thus be made of any of the above models (e.g., by GAM at the national level and XGB at the regional and station levels). For the sake of clarity, the assessment of the probabilistic forecasts are only reported for one base predictor, that consists of Combi at each aggregation level, and one reconciliation procedure, that is rOLS. This choice is motivated by the robustness of this aggregate predictor, and by the modeling rationale of rOLS described in Section 3.1, for the reconciliation procedure. A posterior analysis of the results in point forecasting confirms this choice, as reported in Figure 6. The metric used to assess the point forecast accuracy is the root mean squared error (RMSE), computed for each node $i \in [m]$ over the active test points.

The boxplots report the node-wise RMSEs, where the variability of the RMSEs is assessed across the different node of each level in the 10 experimental runs. GAM exhibits the best

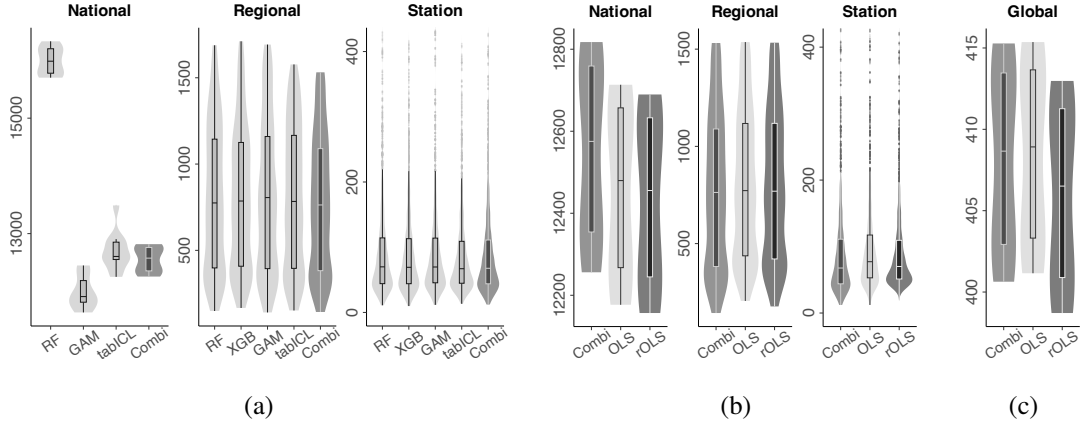


Fig 6: Boxplots of the node-wise RMSEs of the base forecasts (a) and reconciled forecasts (b) across aggregation levels. The Global entry (c) corresponds to the square root of the average MSE over all nodes in the hierarchy.

performance at the national level, Combi at the regional level, and tabICL at the station level (panel (a) in Figure 6). Overall, the performance between the different models and the choice of Combi as the base predictor is essentially driven by its robustness across the hierarchy, as it consistently ranks among the top two models at each level. Note that the random forest is the worst-performing model at the national level by a significant amount. Still, it is included in the Combi predictor, which slightly degrades its overall performance.

Given this choice of base predictor, the panels (b) and (c) in Figure 6 report the RMSEs of the reconciled forecasts respectively at the national, regional, and station levels, and from a global perspective, where the RMSE is computed as the square root of the average MSE over all nodes in the hierarchy. The different reconciliation procedures exhibit similar performance at the regional and station levels, but rOLS is the best-performing procedure both at the national level and from a global perspective. In particular, from a global perspective, enforcing coherence through OLS degrades the performance of the base predictor, while rOLS improves it.

For the sake of completeness, the different reconciliation procedures are also compared in terms of probabilistic forecasts. The analysis, deferred to the Supplementary Material C, confirms the choice of rOLS as the reconciliation procedure, and shows that it yields smaller prediction intervals than both Direct and OLS, while maintaining the desired coverage level.

5.3. Validity rates across all nodes within the hierarchy. This section reports the evaluation of the procedures from a global perspective through validity rates, defined as proportions of nodes at a given level of the hierarchy for which the empirical coverage exceed a specified coverage level. The validity rates are computed either conditional on the node being active or conditional on the node being active and the neighboring nodes being in a certain activity regime, which will be defined in due course.

5.3.1. Validity conditional on the node being active. For each node $i \in [m]$, empirical coverage estimates are built conditional on the node being active. The estimates computed on less than 20 test points are excluded from the validity rate computation. Validity rates are thus computed, for each run of the experiments, over 1 node at the national level, 32 nodes at the regional level, and 1159 of the 1398 nodes at the station level. Given a nominal coverage level $1 - \alpha = 90\%$, the primary objective (depicted in Section 2.3) is to have valid nodes

TABLE 3

Validity rates at two coverage levels: the nominal level $1 - \alpha = 90\%$ and a relaxed level $1 - 2\alpha = 80\%$. Results are grouped by aggregation level/conformal procedure, and reported as mean $\pm 1.96 \times$ standard error.

Node type	Level	CP	CP-MNR	CP-MNR Nested*	Adaptive CP-MNR
National	$1 - \alpha$	40.0% $\pm$ 16.3%	40.0% $\pm$ 16.3%	90.0% $\pm$ 10.0%	0.0% $\pm$ 0.0%
	$1 - 2\alpha$	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
Regional	$1 - \alpha$	51.9% $\pm$ 1.2%	51.9% $\pm$ 1.2%	71.2% $\pm$ 1.2%	47.8% $\pm$ 0.9%
	$1 - 2\alpha$	98.8% $\pm$ 0.5%	98.8% $\pm$ 0.5%	99.1% $\pm$ 0.5%	100.0% $\pm$ 0.0%
Station	$1 - \alpha$	38.1% $\pm$ 0.4%	56.2% $\pm$ 0.4%	73.6% $\pm$ 0.6%	57.6% $\pm$ 0.6%
	$1 - 2\alpha$	67.7% $\pm$ 0.3%	93.6% $\pm$ 0.3%	90.8% $\pm$ 0.4%	98.0% $\pm$ 0.1%

at this coverage level. Hence, higher values for the validity rates generally indicate better calibration. However, a validity rate close to 100% is not necessarily desirable, as it may simply reflect systematic overcoverage due to excessively conservative prediction intervals. As a consequence, one expects a well-performing method to achieve a validity rate at the nominal level above 50% without necessarily approaching 100%. A less ambitious validity criterion is also considered, namely the validity rates at the relaxed level $1 - 2\alpha = 80\%$. The choice to include a relaxed level is motivated by the fact that methods that cover 89% of the test points are not valid at the nominal coverage level $1 - \alpha = 90\%$, but are more reliable than methods that cover only 70% of the test points. The validity rates at the relaxed level thus distinguish methods that are close to the nominal level and those that are not. In that case, a validity rate close to 100% is desirable.

Table 3 reports the validity rates for each conformal procedure and at two coverage levels: the nominal level $1 - \alpha = 90\%$ and a relaxed level $1 - 2\alpha = 80\%$. At the nominal level, CP-MNR-Nested* is the best-performing procedure at each level of the hierarchy by a significant amount, with an average validity rate of 90% at the Scotland node, of 71.2% at the regional level nodes, and of 73.6% at the station level nodes. At the relaxed level, all four procedures are always valid at the national node, and almost always valid at the regional level, with the lowest average validity rate being 98.8% for CP and CP-MNR. At the station level, however, Adaptive CP-MNR outperforms the other procedures, with a validity rate of 98% at the relaxed level, compared to 90.8% for CP-MNR-Nested*. Hence, fewer nodes suffer from a substantial lack of coverage with Adaptive CP-MNR, which is a desirable robustness property from a practical standpoint. A mild undercoverage, say 89% instead of the nominal 90%, has limited operational consequences. A severe one, say 70%, does not: it repeatedly under-estimates the required charging capacity, triggers costly last-minute network reinforcements, and erodes the operator's trust in the forecasts. Since the identity of the poorly covered station is not known in advance, avoiding such worst-case failures anywhere in the network matters more in practice than achieving the nominal level everywhere. The CP and CP-MNR procedures coincide at the national and regional levels, as indicated in Remark 1, so that they exhibit the same validity rates at these levels. Nevertheless, they differ at the station level, where CP-MNR exhibits a substantially higher validity rate than CP (with 56.2% vs 38.1% at the nominal level, and 93.6% vs 67.7% at the relaxed level). This confirms the benefit of accounting for the activity of the node under consideration.

5.3.2. Validity conditional on the activity of the neighboring nodes. In addition to the validity criterion discussed above, that is (only) conditional on the node being active, we are interested in validity conditional on the overall infrastructure activity as motivated in Section 2.3. The station level nodes exhibit the importance of conditional validity for operational decision making since the underlying uncertainty depends on the number of active stations in

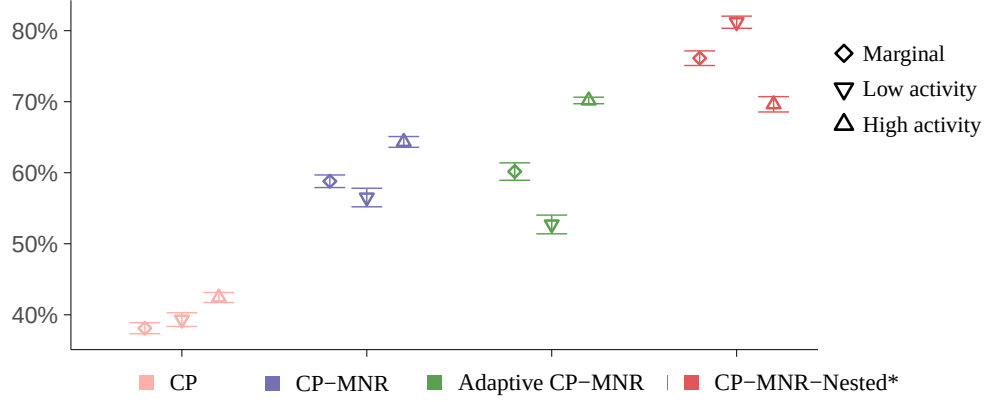


Fig 7: Validity rates at the station-level, respectively marginally and conditionally on the activity regime (Low or High) within the region of the station under scrutiny.

their neighborhood that may affect the accuracy of the forecaster. For each active station, the test period is decomposed into two regimes, depending on the number of active stations in their local authority: a low-activity regime, when the number of active stations is below the median number of active stations over the test period, and a high-activity regime otherwise. This partitioning provides a simple way to demonstrate the conditional validity of the prediction intervals with respect to the activity of the charging infrastructure. Figure 7 reports the validity rates at level $1 - \alpha = 90\%$ respectively marginally and conditional on the low-activity and high-activity regimes. Again, only empirical coverages computed on more than 20 test points are included in the validity rates.

The results indicate that the proportion of valid nodes is systematically lower in the low-activity regime than in the high-activity regime for CP, CP-MNR, and Adaptive CP-MNR. This suggests that the uncertainty around the forecast is higher when the number of neighboring node is low and that these three conformal procedures fail to account for the infrastructure activity. On the contrary, CP-MNR-Nested* demonstrates a similar validity rate than Adaptive CP-MNR in the high-activity regime (both with 70%) but a significantly higher validity rate in the low-activity regime (81% against 53%). It emphasizes that the construction of CP-MNR-Nested* successfully account for the whole infrastructure activity when producing the prediction intervals. The comparison between CP-MNR and Adaptive CP-MNR in Figure 7 also illustrates the fact that the online extension may hinder conditional coverage (as discussed in Section 4.3). Indeed, the gap in validity rates between the low-activity and high-activity regimes is widened by the adaptive extension from 57%–64% to 52%–70%.

The numerical results obtained so far indicate that CP-MNR-Nested* provides the best overall performance in terms of the two validity measures. We now examine whether these empirical guarantees translate into prediction intervals that are also meaningful from a practical standpoint. To this end, prediction intervals are displayed for a selection of representative nodes from the hierarchy. Besides illustrating the methods, these examples provide a node-level assessment of the coverage-length trade-off and reveal how each method adapts to different infrastructure activity patterns.

5.4. Focus on Scotland, Glasgow, and two stations in Glasgow. The Scotland node, depicted in Figure 2, corresponds to the highest aggregation level of the hierarchy. Its demand profile exhibits a pronounced downward trend over the study period, mainly driven by the gradual introduction of charging tariffs across Scotland. As an aggregated node, it remains

only weakly affected by changes in the activity of individual charging stations. The Glasgow node, also depicted in Figure 2, corresponds to a large urban area with a mature yet continuously evolving charging infrastructure. At the station level, Stations 406 and 386 are two representative nodes in Glasgow, displayed in the right-hand panel of Figure 3. Station 406 became operational only a few months before the beginning of the test period, whereas Station 386 reappeared halfway through the test period after a prolonged period of inactivity following a brief operational phase in 2022. These two station-level nodes are expected to constitute more challenging forecasting tasks, especially Station 386, since the calibration set may contain only a small number of observations for which the stations are active.

5.4.1. Node-wise coverages and lengths. Figure 8 reports the empirical coverage and average length for the four nodes under scrutiny. A procedure that is valid at the nominal level $1 - \alpha = 90\%$ is on the right of the dashed line in Figure 8, and the efficiency of a procedure is granted by a shorter average length of the prediction intervals, i.e., located at the bottom of the plots. The prediction intervals produced by the CP-MNR and Adaptive CP-MNR procedures exhibit a satisfying coverage at all nodes, with a minimal empirical coverage of 84% for CP-MNR and 88% for Adaptive CP-MNR, both for Station 406. It confirms the findings of Table 3 obtained with validity rates at the relaxed coverage level: the above procedures are closely aligned with the nominal coverage level of 90%, even if not necessarily valid. The only procedure that is valid for all four nodes is CP-MNR-Nested*, with a coverage of 91% for Scotland, of 97% for Glasgow, and of 91% both for Station 406 and Station 386. Again, it corroborates the fact that CP-MNR-Nested* provides the best validity guarantees at all levels of the hierarchy. The CP procedure, on the other hand, suffers from a significant lack of coverage at the station level, with an empirical coverage of 70% for Station 406 and of (only) 9% for Station 386. This highlights the importance of accounting for the activity of the node under consideration when constructing the prediction intervals, as done by the CP-MNR, Adaptive CP-MNR, and CP-MNR-Nested*. In terms of efficiency, the Adaptive CP-MNR

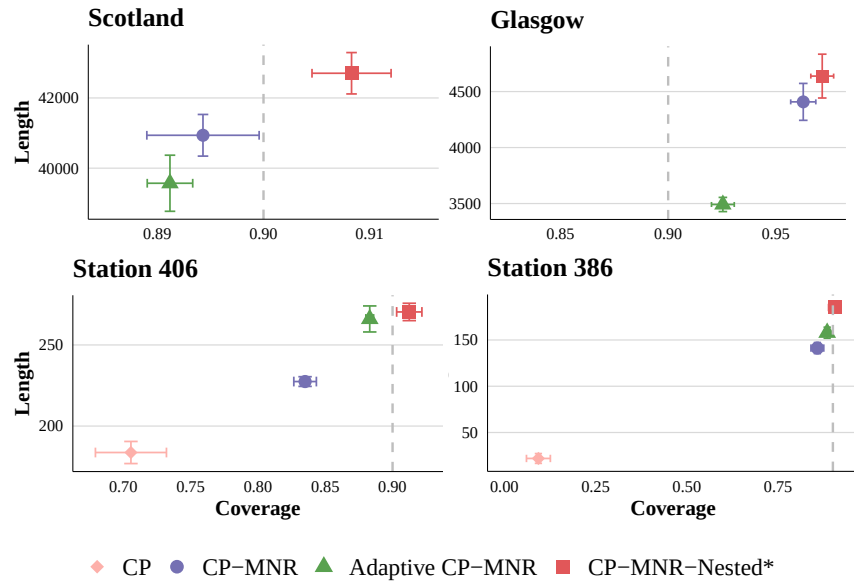


Fig 8: Coverage and average length of the prediction intervals for Scotland (top left), Glasgow (top right), Station 406 (bottom left), and Station 386 (bottom right). The results are reported as $\text{mean} \pm 1.96 \times \text{standard error}$.

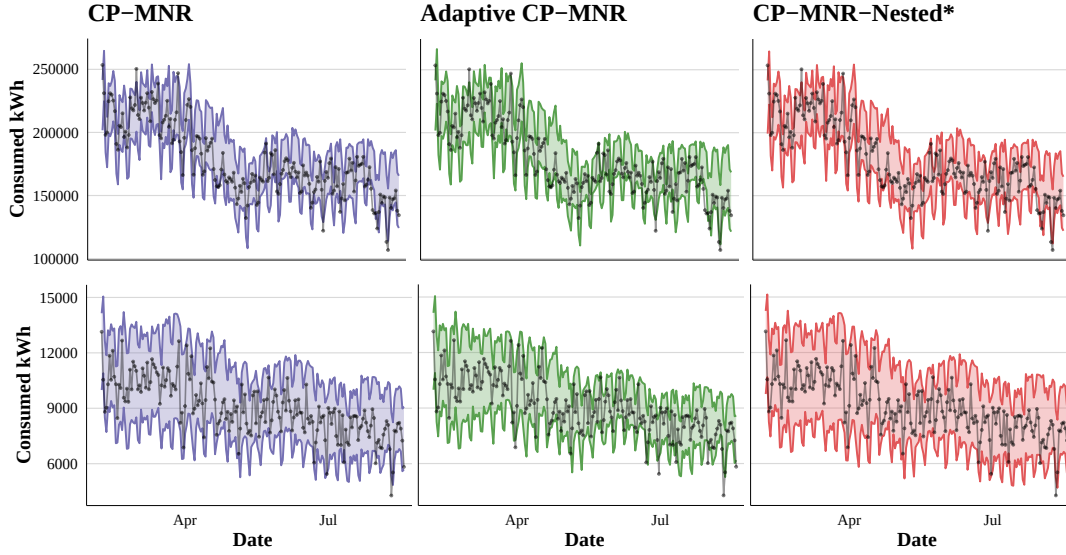


Fig 9: Prediction intervals for Scotland (top) and Glasgow (bottom).

procedure yields significantly smaller prediction intervals in the case of Glasgow (a relative improvement of at least 20%), while being valid, suggesting that Adaptive CP-MNR is the best compromise between validity and efficiency. By contrast, CP-MNR-Nested* produces the largest prediction intervals on average, for all four nodes, which emphasizes the trade-off between validity and efficiency.

5.4.2. Visual assessment of the prediction intervals. To complement the numerical results, the performances are visually assessed by reporting in Figures 9–10–11 the prediction intervals output by all procedures for one run of the experiment.

At the aggregated levels, namely in Figure 9, the visual inspection of the prediction intervals indicates that the procedures are able to handle the temporal evolution of the data. Adaptive CP-MNR is designed to adapt to the temporal evolution of the data but the fact that CP-MNR and CP-MNR-Nested* also exhibit this behavior suggests that updating the underlying reconciled predictors every two weeks is sufficient to capture this evolution. The prediction intervals produced by CP-MNR and CP-MNR-Nested* seem overly conservative in the case of Glasgow, especially at the end of the test period, whereas Adaptive CP-MNR adjusts to this accurately predicted period, and produces smaller prediction intervals, while remaining valid. This illustrates the benefit of the adaptive extension, that progressively reduces the size of the prediction intervals if the coverage level is above the nominal level.

At the station level, Station 406 (Figure 10) and Station 386 (Figure 11) both illustrate two typical traits of the prediction intervals. The first attribute is that the CP procedure produces prediction intervals that are systematically too narrow, leading to a significant undercoverage of the test points, whereas the other three procedures produce prediction intervals that are close to the nominal coverage level in terms of empirical coverage. This is particularly visible in Figure 11, where the energy demand is zero for most of the calibration points. As a consequence, the CP-MNR, Adaptive CP-MNR, and CP-MNR-Nested* procedures does not produce prediction intervals until some non-zero calibration points are observed (this is why only CP output prediction intervals for the first three non-zero observations of the test period). The same phenomenon is observed for Station 406, but to a smaller extent: the proportion of zero energy demands is lower than for Station 386, so that they have a smaller impact on the prediction intervals produced by CP.

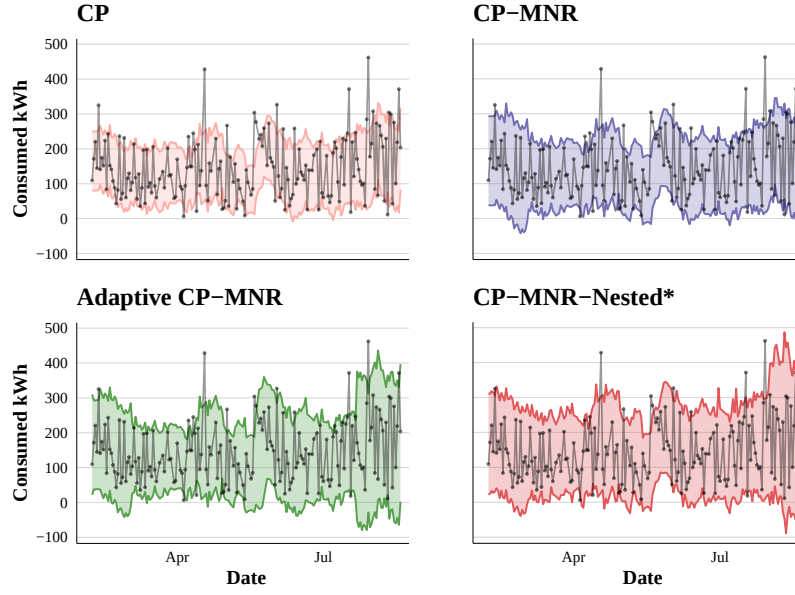


Fig 10: Prediction intervals for Station 406.

The second attribute arises at the end of the test period and is related to the activity of neighboring stations. On August 10th and 11th, the number of active stations in Glasgow is significantly lower than usual, namely 45 and 46 compared to 67 active stations on average over the test period. At these two dates, CP-MNR-Nested* produces significantly larger prediction intervals than it does for the rest of the test period, which is not the case for the other procedures. This observation suggests that the CP-MNR-Nested* procedure accounts for the charging infrastructure activity in a more sensitive way than the other procedures, since it detects that the uncertainty is higher when the number of active stations is lower than usual.

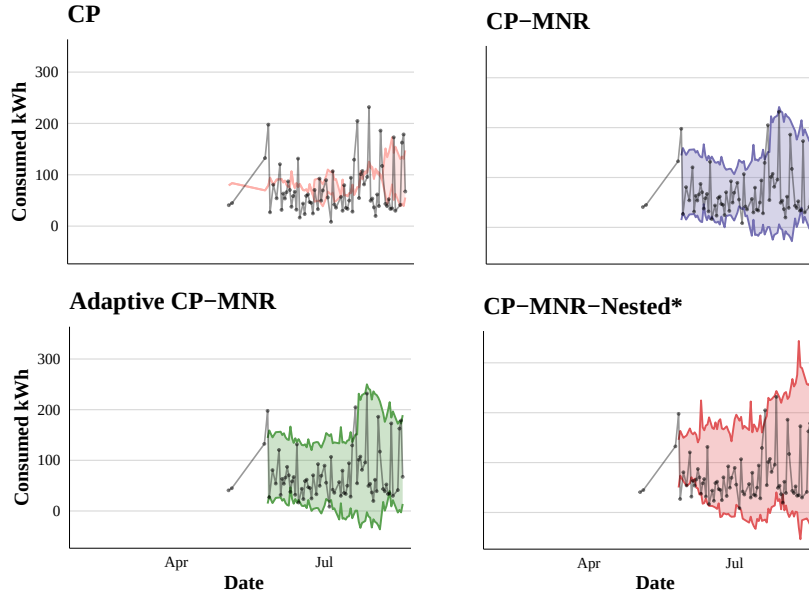


Fig 11: Prediction intervals for Station 386.

The above phenomenon are consistent with the findings of Figure 7, where every other procedure exhibits a lower validity rate in the low-activity regime than in the high-activity regime. Note that the Adaptive CP-MNR procedure also produces larger prediction intervals at the end of the test period, but this behavior is driven by the fact that it fails to cover two consecutive at the end of July, rather than by some adaptive consideration of the activity of the charging infrastructure.

6. Discussion and future work. We successfully combined forecast reconciliation and conformal prediction in the time-varying hierarchical structure of the EV charging network from the ChargePlace Scotland data. The resulting forecasts and prediction intervals are valuable assets (e.g., as entries of an optimization algorithm) that accurately reflect the uncertainty of the forecasts for the current state of the charging infrastructure. The numerical experiment demonstrates that CP-MNR-Nested* produces valid prediction intervals for different configurations of the charging infrastructure, whereas CP-MNR and Adaptive CP-MNR, yield prediction intervals that are valid conditionally on the activity of the node under consideration, but that are less sensitive to the activity of the whole charging infrastructure.

A natural direction for future work is to increase the adaptivity of the methodology through online extensions of the learning and reconciliation procedures. In this direction, one can build on the methodology of Br  g  re and Huard (2022), that combines aggregation of experts with OLS reconciliation, and of Di Modica, Pinson and Taieb (2021); Hihat et al. (2024) that propose online multivariate reconciliation procedures based respectively on a least-squares and a regularized least-squares online estimation.

Acknowledgments. The authors would like to thank Margaux Zaffran for the fruitful discussions on her work on conformal prediction for missing data.

Funding. This work was conducted as part of the Inria-EDF challenge.

SUPPLEMENTARY MATERIAL

Connection with CP-MDA and Theoretical Results

This supplement gives mathematical insights that motivate the construction of CP-MNR and CP-MNR-Nested*. It includes the conceptual connection between our conformal prediction with missing node reconciliation (MNR) setting and the conformal prediction with missing data (MDA) framework developed by Zaffran et al. (2024) that guided our approach.

Implementation of the Algorithms

This supplement provides further information on the implementation of the proposed algorithms. It includes (i) the formal description of the learning and reconciliation algorithms that are used in Algorithms 1 and 2; (ii) the implementation of the conformal step in Algorithm 2; (iii) the implementation and theoretical analysis of the Adaptive CP-MNR procedure introduced in Section 4.3.

Additional Numerical Results

This supplement focuses on the comparison of the different forecast reconciliation procedures in terms of the prediction interval coverage and length. It advocates for the use of rOLS, which is the focus of the main text.

REFERENCES

- ALARAJ, M., RADI, M., ALSISI, E., MAJDALAWIEH, M. and DARWISH, M. (2025). Machine Learning-Based Electric Vehicle Charging Demand Forecasting: A Systematized Literature Review. *Energies* **18** 4779.
- AMARA-OUALI, Y., HAMROUCHE, B., PRINCIPATO, G. and GOUDE, Y. (2025). Quantifying the uncertainty of electric vehicle charging with probabilistic load forecasting. *World Electric Vehicle Journal* **16** 88.
- ANGELOPOULOS, A. N. and BATES, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning* **16** 494–591.
- ANTONIADIS, A., CUGLIARI, J., FASIOLO, M., GOUDE, Y. and POGGI, J.-M. (2024). *Statistical Learning Tools for Electricity Load Forecasting*. Springer.
- ATHANASOPOULOS, G., AHMED, R. A. and HYNDMAN, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* **25** 146–166.
- BARBER, R. F., CANDES, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics* **49** 486–507.
- BATIC, D. (2025). GridCharge: Capturing Carbon Intensity of 5M+ Public EV Charging Sessions in Scotland. <https://huggingface.co/datasets/djordjebatic/GridCharge>.
- BOGDANOV, D. and BREYER, C. (2024). Role of smart charging of electric vehicles and vehicle-to-grid in integrated renewables-based energy systems on country level. *Energy* **301** 131635.
- BOJER, C. S. and MELDGAARD, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting* **37** 587–603.
- BRÉGÈRE, M. and HUARD, M. (2022). Online hierarchical forecasting for power consumption data. *International Journal of Forecasting* **38** 339–351.
- BREIMAN, L. (2001). Random forests. *Machine learning* **45** 5–32.
- CHALLU, C., OLIVARES, K. G., ORESHKIN, B. N., RAMIREZ, F. G., CANSECO, M. M. and DUBRAWSKI, A. (2023). Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* **37** 6989–6997.
- CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I., ZHOU, T. et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2* **1** 1–4.
- DAVIES, K., HART, E. and GALLOWAY, S. (2024). Plugging-in Caledonia: location and utilisation of public electric vehicle chargers in Scotland. *World Electric Vehicle Journal* **15** 570.
- DE VILMAREST, J., BROWELL, J., FASIOLO, M., GOUDE, Y. and WINTENBERGER, O. (2023). Adaptive probabilistic forecasting of electricity (net-) load. *IEEE Transactions on Power Systems* **39** 4154–4163.
- DI MODICA, C., PINSON, P. and TAIEB, S. B. (2021). Online forecast reconciliation in wind power prediction. *Electric Power Systems Research* **190** 106637.
- DIAZ-LONDONO, C., MAFFEZZONI, P., DANIEL, L. and GRUOSSO, G. (2024). Comparison and Analysis of Algorithms for Coordinated EV Charging to Reduce Power Grid Impact. *IEEE Open Journal of Vehicular Technology* **5** 990–1003.
- ERICKSON, N., PURUCKER, L., TSCHALZEV, A., HOLZMÜLLER, D., DESAI, P., SALINAS, D. and HUTTER, F. (2026). Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems* **38**.
- FESCIOGLU-UNVER, N. and AKTAŞ, M. (2023). Electric vehicle charging service operations: A review of machine learning applications for infrastructure planning, control, pricing and routing. *Renewable and Sustainable Energy Reviews* **188** 113873.
- GIBBS, I. and CANDÈS, E. J. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems* **34** 1660–1672.
- GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1** 125–151.
- GOUDE, Y., NEDELLEC, R. and KONG, N. (2013). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE transactions on smart grid* **5** 440–446.
- GUPTA, C., KUCHIBHOTLA, A. K. and RAMDAS, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition* **127** 108496.
- HIHAT, M., GARRIGOS, G. and BUSSY, S. (2024). Multivariate online linear regression for hierarchical forecasting. *arXiv preprint arXiv:2402.14578*.
- HOLLMANN, N., MÜLLER, S., PURUCKER, L., KRISHNAKUMAR, A., KÖRFER, M., HOO, S. B., SCHIRRMESTER, R. T. and HUTTER, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature* **637** 319–326.
- HOLLYMAN, R., PETROPOULOS, F. and TIPPING, M. E. (2021). Understanding forecast reconciliation. *European Journal of Operational Research* **294** 149–160.
- HYNDMAN, R. J., AHMED, R. A., ATHANASOPOULOS, G. and SHANG, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* **55** 2579–2589.

- IEA (2022). Electric Vehicles. <https://www.iea.org/reports/electric-vehicles>. Accessed: 2026-04-21.
- KHAN, W., SOMERS, W., WALKER, S., DE BONT, K., VAN DER VELDEN, J. and ZEILER, W. (2023). Comparison of electric vehicle load forecasting across different spatial levels with incorporated uncertainty estimation. *Energy* **283** 129213.
- KIM, Y. and KIM, S. (2021). Forecasting charging demand of electric vehicles using time-series models. *Energies* **14** 1487.
- LANDAU, H. G. (1953). On dominance relations and the structure of animal societies. III. The condition for a score structure. *Bull. Math. Biophys* **15** 143–148.
- LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113** 1094–1111.
- MAKRIDAKIS, S., PETROPOULOS, F. and SPILOTIS, E. (2022). The M5 competition: Conclusions. *International Journal of Forecasting* **38** 1576–1582. Special Issue: M5 competition.
- MURATORI, M. (2018). Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nature Energy* **3** 193–201.
- NUTKANI, I., TOOLE, H., FERNANDO, N. and ANDREW, L. P. C. (2024). Impact of EV charging on electrical distribution network and mitigating solutions—A review. *IET Smart Grid* **7** 485–502.
- OSTERMANN, A. and HAUG, T. (2024). Probabilistic forecast of electric vehicle charging demand: analysis of different aggregation levels and energy procurement. *Energy Informatics* **7** 13.
- PANAGIOTELIS, A., ATHANASOPOULOS, G., GAMAKUMARA, P. and HYNDMAN, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting* **37** 343–359.
- PIERROT, A. and GOUDE, Y. (2011). Short-term electricity load forecasting with generalized additive models In *16th Intelligent System Applications to Power Systems Conference*.
- POWELL, S., CEZAR, G. V., MIN, L., AZEVEDO, I. M. L. and RAJAGOPAL, R. (2022). Charging infrastructure access and operation to reduce the grid impacts of deep electric vehicle adoption. *Nature Energy* **7** 932–945.
- PRINCIPATO, G., STOLTZ, G., AMARA-OUALI, Y., GOUDE, Y., HAMROUCHE, B. and POGGI, J.-M. (2026). Conformal prediction for hierarchical data. *Transaction on Machine Learning Research*.
- QU, J., HOLZMÜLLER, D., VAROQUAUX, G. and LE MORVAN, M. (2025). TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *International Conference on Machine Learning* 50817–50847. PMLR.
- ROMANO, Y., BARBER, R. F., SABATTI, C. and CANDÈS, E. (2020). With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review* **2** 4.
- SALMON, M. and ELCHISON, J. (2025). riem: Accesses Weather Data from the Iowa Environment Mesonet R package version 1.0.0.
- SHCHUR, O., ANSARI, A. F., TURKMEN, C., STELLA, L., ERICKSON, N., GUERRON, P., BOHLKE-SCHNEIDER, M. and WANG, Y. (2025). fev-bench: A realistic benchmark for time series forecasting. Preprint arXiv:2509.26468.
- VAN ETEN, T., DEGELER, V. and LUO, D. (2024). Large-scale forecasting of electric vehicle charging demand using global time series modeling. In *International Conference on Vehicle Technology and Intelligent Transport Systems* 40–51. SciTePress.
- VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer.
- ŠTOGL, O., MILTNER, M., ZANOCCO, C., TRAVERSO, M. and STARÝ, O. (2024). Electric vehicles as facilitators of grid stability and flexibility: A multidisciplinary overview. *WIREs Energy and Environment* **13** e536.
- WANG, S., LI, Y., SHAO, C., WANG, P., WANG, A. and ZHUGE, C. (2025). An adaptive spatio-temporal graph recurrent network for short-term electric vehicle charging demand prediction. *Applied Energy* **383** 125320.
- WICKRAMASURIYA, S. L., ATHANASOPOULOS, G. and HYNDMAN, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association* **114** 804–819.
- WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall.
- YI, Z., LIU, X. C., WEI, R., CHEN, X. and DAI, J. (2022). Electric vehicle charging demand forecasting using deep learning model. *Journal of Intelligent Transportation Systems* **26** 690–703.
- ZAFFRAN, M., FÉRON, O., GOUDE, Y., JOSSE, J. and DIEULEVEUT, A. (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning* 25834–25866. PMLR.
- ZAFFRAN, M., DIEULEVEUT, A., JOSSE, J. and ROMANO, Y. (2023). Conformal prediction with missing values. In *International Conference on Machine Learning* 40578–40604. PMLR.
- ZAFFRAN, M., JOSSE, J., ROMANO, Y. and DIEULEVEUT, A. (2024). Predictive uncertainty quantification with missing covariates. Preprint arXiv:2405.15641.
- ZHANG, J., WANG, Z., MILLER, E. J., CUI, D., LIU, P. and ZHANG, Z. (2023). Charging demand prediction in Beijing based on real-world electric vehicle data. *Journal of Energy Storage* **57** 106294.
- ZHENG, K., XU, H., LONG, Z., WANG, Y. and CHEN, Q. (2023). Coherent hierarchical probabilistic forecasting of electric vehicle charging demand. *IEEE Transactions on Industry Applications* **61** 1329–1340.

A. CONNECTION WITH CP-MDA AND THEORETICAL RESULTS

This Supplementary Material presents the heuristic theoretical coverage guarantees of the CP-MNR and CP-MNR-Nested* procedures. The analysis is based on the connection between the CP-MNR setting and the conformal prediction with missing data framework of Zaffran et al. (2023, 2024), which guided the approach. We first establish this connection.

A.1. Connection with the conformal prediction with missing data framework. The incentive for the CP-MNR and CP-MNR-Nested* procedures is to construct prediction intervals that are valid conditionally on the state of the infrastructure, represented by a binary activity vector. This is closely related to the objective of the missing data augmentation (MDA) framework of Zaffran et al. (2023, 2024), where validity is conditioned on a binary mask indicating which features are missing. Nevertheless, the binary vector of infrastructure activity does not encode missingness but instead identifies inactive nodes in the hierarchy, whose zero-valued constraints directly affect the forecast of every node of the hierarchy, through the reconciliation procedure.

The two settings are thus conceptually different, but the following observation establishes a connection between them:

“Base forecasts are covariates in the reconciliation step, and enforcing the infrastructure activity amounts to setting some base forecasts to zero, which corresponds to a specific imputation strategy in the framework of conformal prediction with missing data.”

This observation explicit is made explicit by building on the interpretation from Hollyman, Petropoulos and Tipping (2021) of forecast reconciliation as forecast combination. This interpretation is illustrated on a simple example, omitting algebraic details for brevity.

Consider the toy-example used throughout the main text in Sections 2.2 and 3.1, with $n = 3$ stations, $m = 4$ nodes, infrastructure activity vector $\mathbf{a} = (1, 0, 1, 1)^\top$, and structural matrix H defined in (3). Let $\hat{\mathbf{y}}$ denote a base forecasts vector and denote by $\tilde{\mathbf{y}}$ the associated reconciled forecasts vector obtained by applying a projection P onto the image of H to the base forecasts vector. Note that P can be the orthogonal projection as in the OLS reconciliation procedure, but it can also be any other projection onto the image of H .

Following Hollyman, Petropoulos and Tipping (2021), the reconciled forecasts $\tilde{\mathbf{y}}$ can be expressed as a constrained linear combination of direct forecast of the node under scrutiny, and an indirect one, constructed from the remaining nodes:

$$\begin{aligned}\tilde{y}_4 &= (3 - \lambda_1 - \lambda_2 - \lambda_3) \hat{y}_4 + (\lambda_1 + \lambda_2 + \lambda_3 - 2)(\hat{y}_1 + \hat{y}_2 + \hat{y}_3); \\ \tilde{y}_1 &= \lambda_1 \hat{y}_1 + (1 - \lambda_1)(\hat{y}_4 - \hat{y}_2 - \hat{y}_3); \quad \tilde{y}_2 = \lambda_2 \hat{y}_2 + (1 - \lambda_2)(\hat{y}_4 - \hat{y}_1 - \hat{y}_3); \\ \tilde{y}_3 &= \lambda_3 \hat{y}_3 + (1 - \lambda_3)(\hat{y}_4 - \hat{y}_1 - \hat{y}_2), \quad \text{where } \lambda_1, \lambda_2, \lambda_3 \text{ depend on the projection } P.\end{aligned}$$

Now, the reconciled forecasts $\tilde{\mathbf{y}}'$ produced by rOLS can be expressed in a similar way:

$$\begin{aligned}\tilde{y}'_4 &= (2 - \lambda'_1 - \lambda'_3) \hat{y}_4 + (\lambda'_1 + \lambda'_3 - 1)(\hat{y}_1 + 0 + \hat{y}_3); \\ \tilde{y}'_1 &= \lambda'_1 \hat{y}_1 + (1 - \lambda'_1)(\hat{y}_4 - 0 - \hat{y}_3); \quad \tilde{y}'_2 = 0; \quad \tilde{y}'_3 = \lambda'_3 \hat{y}_3 + (1 - \lambda'_3)(\hat{y}_4 - \hat{y}_1 - 0),\end{aligned}$$

with $\lambda'_1 = \lambda'_3 = 2/3$ in this case.

This formulation makes it clear that the base forecasts of inactive nodes no longer contribute to the forecast combination: $\hat{y}_2$ is systematically replaced by zero. This mechanism can be interpreted as a particular imputation strategy, where the base forecasts of the inactive nodes are deterministically imputed by zero before constructing the indirect forecasts.

This perspective reveals a close connection with the conformal prediction with missing data framework of Zaffran et al. (2024), allowing the theoretical analysis to follow essentially the same arguments.

A.2. Theoretical Result for CP-MNR. In this section, we present the theoretical coverage guarantee for CP-MNR, which is conditional on the activity of the node under scrutiny.

Following the procedure described in Sections 3.2 and 4, we consider a calibration set of observations indexed by Cal with cardinality $\#\text{Cal}$, and a test point $(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$ with unknown target value $\mathbf{y}^{(t)}$. Recall that the non-conformity score function introduced in (7) and considered throughout the theoretical analysis is defined as

$$(12) \quad \forall i \in [m], \quad s_i : ((\mathbf{x}, \mathbf{a}), \mathbf{y}; \tilde{A}) \mapsto \left| \Pi_i(\mathbf{y} - \tilde{A}(\mathbf{x}, \mathbf{a})) \right|.$$

As in classic conformal prediction, the theoretical framework is built on an exchangeability assumption. The latter may not hold in practice, but provides a heuristic justification for the validity of the proposed procedures.

ASSUMPTION 1. The random variables $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal} \cup \{t\}}$ are exchangeable, i.e., their joint distribution is invariant under permutations.

The following theorem provides a theoretical coverage guarantee for CP-MNR, that is conditional on the activity of the node under scrutiny.

THEOREM A.1. *Consider the nominal coverage level $1 - \alpha \in [0, 1]$. Under Assumption 1, the CP-MNR procedure of Algorithm 1, run with any learning and reconciliation algorithms $\mathcal{A}, \mathcal{R}$, satisfies*

$$(13) \quad \forall i \in [m], \quad \mathbb{P}\left(y_i^{(t)} \in \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) \mid a_i^{(t)} = 1\right) \geq 1 - \alpha.$$

PROOF. The result follows from the fact that Algorithm 1 outputs infrastructure activity-conditional Mondrian conformal predictors (Vovk, Gammerman and Shafer, 2005). Precisely, our formulation corresponds to the group-conditional framework of Romano et al. (2020). For the sake of completeness, we report their arguments.

Fix $i \in [m]$. Conditional on $a_i^{(t)} = 1$, the grouping of calibration indices in Algorithm 1 is of the form $\mathcal{I}_1 = \{k \in \text{Cal} : a_i^{(k)} = 1\}$. Thus, the exchangeability property of Assumption 1 implies that $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)})\}_{k \in \mathcal{I}_1 \cup \{t\}}$ is exchangeable conditional on $a_i^{(t)} = 1$.

The reconciled predictor $\tilde{A}$ only depends on data from the training set, so that the non-conformity scores $\{s_i^{(k)}\}_{k \in \mathcal{I}_1 \cup \{t\}}$ is exchangeable conditional on $a_i^{(t)} = 1$.

Finally, resorting to Romano et al., 2020, Lemma 1, which is essentially the same argument as in the classical conformal prediction setting adapted to conditional exchangeability, we obtain the infrastructure activity-conditional guarantee in (13), as announced. $\square$

REMARK 3. The CP-MNR procedure of Algorithm 1 has no direct counterpart in Zaffran et al. (2024). For instance, in MDA, there is no particular reason to assess validity conditional on missingness of one specific covariate rather than another. By contrast, in MNR, the activity of the node under consideration is a particularly meaningful conditioning variable, as it directly shapes the demand behavior of the node.

A.3. Theoretical Results for CP-MNR-Nested*. We show two theoretical coverage guarantees for the CP-MNR-Nested* procedure of Algorithm 2. The first one is a marginal guarantee, obtained under the mild exchangeability assumption of Assumption 1. The second one is a conditional guarantee, obtained under a stronger assumption, as described next.

The first result is a weaker result than the one of Theorem A.1 in two ways: it is marginal rather than locally conditional, and the coverage level is $1 - 2\alpha$ rather than $1 - \alpha$.

THEOREM A.2. *Consider the nominal coverage level $1 - \alpha \in [0, 1]$. Under Assumption 1, the CP-MNR-Nested* procedure of Algorithm 2, run with $\widetilde{\text{Cal}} = \text{Cal}$, any learning and reconciliation algorithms $\mathcal{A}$, $\mathcal{R}$, satisfies*

$$(14) \quad \forall i \in [m], \quad \mathbb{P}\left(y_i^{(t)} \in \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})\right) \geq 1 - 2\alpha.$$

PROOF. The proof follows the lines and arguments from the proof of Zaffran et al., 2024, Theorem 4.2, that is itself inspired from the jackknife+ conformal predictors proof from Barber et al., 2021, Theorem 1. For the sake of completeness, we report their proofs to our setting and notations.

Fix $i \in [m]$, and denote by $\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$ the corresponding conformal prediction set output by Algorithm 2, when run with $\widetilde{\text{Cal}} = \text{Cal}$. By construction, we have that

$$(15) \quad \left\{y_i^{(t)} \notin \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})\right\} = \left\{(1 - \alpha)(\#\text{Cal} + 1) \leq \sum_{k \in \text{Cal}} \mathbf{1}_{S_i^{(k),t} < S_i^{(t),k}}\right\},$$

where $S_i^{(p),q}$ denotes the non-conformity score of a calibration point p when the infrastructure activity pattern of the point q is applied on top of its own infrastructure activity pattern, for all $p, q \in \text{Cal} \cup \{t\}$.

Following Barber et al. (2021), we define a comparison matrix M_i , indexed by $\text{Cal} \cup \{t\}$, with entries $M_i(p, q) = \mathbf{1}_{S_i^{(p),q} > S_i^{(q),p}}$ for $p, q \in \text{Cal} \cup \{t\}$, and consider the set

$$(16) \quad W(M_i) = \left\{p \in \text{Cal} \cup \{t\} : \sum_{q \in \text{Cal} \cup \{t\}} M_i(p, q) \geq (1 - \alpha)(\#\text{Cal} + 1)\right\}.$$

Thereof, we can rewrite (15) as $\{y_i^{(t)} \notin \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})\} = \{t \in W(M_i)\}$, so that it suffices to upper bound the probability of the event $\{t \in W(M_i)\}$.

Denote by $\mathcal{M}_i$ the function that maps $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal} \cup \{t\}}$ to the comparison matrix M_i defined above, and fix $p, q \in \text{Cal} \cup \{t\}$. For any permutation σ on $\text{Cal} \cup \{t\}$, the symmetrical construction of the reconciled forecast, and thus of the scores, ensures that

$$(17) \quad \begin{aligned} M_i(\sigma(p), \sigma(q)) &= \mathcal{M}_i\left((\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)})_{k \in \text{Cal} \cup \{t\}}\right)(\sigma(p), \sigma(q)) \\ &= \mathcal{M}_i\left((\mathbf{x}^{(\sigma(k))}, \mathbf{a}^{(\sigma(k))}, \mathbf{y}^{(\sigma(k))})_{k \in \text{Cal} \cup \{t\}}\right)(p, q). \end{aligned}$$

Therefore, we combine (17) with Assumption 1, and get that $M_i^\sigma \stackrel{d}{=} M_i$, where M_i^σ is the permuted matrix $(M_i(\sigma(p), \sigma(q)))_{p, q \in \text{Cal} \cup \{t\}}$. As a consequence, $\mathbb{P}(p \in W(M_i))$ does not depend on the specific value of $p \in \text{Cal} \cup \{t\}$, which implies that

$$\mathbb{P}(t \in W(M_i)) = \frac{1}{\#\text{Cal} + 1} \sum_{p \in \text{Cal} \cup \{t\}} \mathbb{P}(p \in W(M_i)) = \frac{1}{\#\text{Cal} + 1} \mathbb{E}[|W(M_i)|].$$

The proof concludes by using a deterministic upper bound on the cardinality of $W(M_i)$, obtained by Landau's theorem for tournaments (Landau, 1953), namely,

$$\#W(M_i) \leq 2\alpha(\#\text{Cal} + 1).$$

□

We now turn to the stronger conditional coverage guarantee of CP-MNR-Nested*, obtained under the following stronger assumptions.

ASSUMPTION 2. The random variables $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal} \cup \{t\}}$ are drawn i.i.d.

ASSUMPTION 3. We assume that the infrastructure activity pattern is independent of the observations, i.e., for all $k \in \text{Cal} \cup \{t\}$, we have that $\mathbf{a}^{(k)} \perp (\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$.

Assumption 2 is a slightly stronger version of Assumption 1, but the main difference is that Assumption 3 is a very restrictive assumption, which does not hold in our setting. Indeed, the infrastructure activity vector is deterministically induced by the EV charging energy demand vector, which breaks the independence assumption of Assumption 3. Nevertheless, the encouraging simulation evidence from Zaffran et al., 2024, Section 5.2 and 5.3, together with our empirical findings in Section 5, indicate that the above Theorem A.3 and the associated conditional guarantee of CP-MNR-Nested* may still hold in practice, even when Assumption 3 is violated.

THEOREM A.3. Consider the nominal coverage level $1 - \alpha \in [0, 1]$. Under Assumption 2 and 3, the CP-MNR-Nested* procedure of Algorithm 2, run with $\bar{\text{Cal}} = \text{Cal}$, any learning and reconciliation algorithms $\mathcal{A}, \mathcal{R}$, satisfies

$$(18) \quad \forall \mathbf{a} \in \{0, 1\}^m, i \in [m], \quad \mathbb{P}\left(y_i^{(t)} \in C_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) \mid \mathbf{a}^{(t)} = \mathbf{a}\right) \geq 1 - 2\alpha.$$

PROOF. As for the proof of Theorem A.2, the proof essentially follows the lines and arguments of Zaffran et al. (2024), namely of Theorem 4.3. We only report our own formulation of their proof with our setting and notations.

Step 1: conditional exchangeability. By Assumptions 2 and 3, the test infrastructure activity pattern is independent from $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$ for all $k \in \text{Cal} \cup \{t\}$ so that the set of observations with substituted infrastructure activity $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(t)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal}}, (\mathbf{x}^{(t)}, \mathbf{a}^{(t)}, \mathbf{y}^{(t)})\}$ is exchangeable conditional on $\mathbf{a}^{(t)}$.

Step 2: from Algorithm 2 to Algorithm 3. We recast CP-MNR-Nested* as a procedure that performs the conformal step on the observations with substituted infrastructure activity patterns, for which we have conditional exchangeability by Step 1. To do so, we introduce Algorithm 3, and show that the CP-MNR-Nested* procedure corresponds to a particular instance of Algorithm 3 run with calibration inputs:

$$\{(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{z}^{(k)})\}_{k \in \text{Cal}} = \{(\mathbf{x}^{(k)}, \mathbf{a}^{(t)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal}}.$$

First, we use that $\tilde{A}(\cdot, \cdot) = \hat{R}(\hat{A}(\cdot, \cdot), \cdot)$ so that, for any observation $(\mathbf{u}, \mathbf{v}, \mathbf{z})$, its non-conformity score at node $i \in [m]$ with augmented infrastructure activity $\tilde{\mathbf{a}} = \min(\mathbf{v}, \mathbf{a}^{(t)})$ of Algorithm 2 rewrites as

$$s_i\left((\mathbf{u}, \tilde{\mathbf{a}}), \mathbf{z}; \hat{R}(\hat{A}(\cdot, \cdot), \cdot)\right) = s_i\left((\mathbf{u}, \mathbf{a}^{(t)}), \mathbf{z}; \hat{R}^*(\hat{A}^*(\cdot, \cdot; \mathbf{v}), \cdot; \mathbf{v})\right),$$

Algorithm 3 CP-MNR-Nested* at level $1 - \alpha$ with random infrastructure activity patterns

Input: Predictor $\hat{R}^*(\hat{A}^*(\cdot, \cdot; \cdot), \cdot; \cdot)$; dataset $\{(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{z}^{(k)}) : k \in \text{Cal}\}$; test point $(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$.

1: **for** $k \in \text{Cal}$ **do** consider a random augmented nested infrastructure activity pattern
 Randomly draw $\mathbf{a}^{(k)}$ independently from $\{(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{z}^{(k)})_{k \in \text{Cal}}, (\mathbf{x}^{(t)}, \mathbf{a}^{(t)}, \mathbf{y}^{(t)})\}$
 Define accordingly a modified reconciled predictor, $\tilde{g}^{(k)}(\cdot, \cdot) = \hat{R}^*(\hat{A}^*(\cdot, \cdot; \mathbf{a}^{(k)}), \cdot; \mathbf{a}^{(k)})$

end for

3: **for** $i \in [m]$ **do** build the prediction set $\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$ defined as

$$\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) = \left\{ \Pi_i(\mathbf{y}) \in \mathbb{R} : (1 - \alpha)(\#\text{Cal} + 1) > \# \left\{ k \in \text{Cal} : s_i((\mathbf{u}^{(k)}, \mathbf{a}^{(t)}), \mathbf{z}^{(k)}; \tilde{g}^{(k)}(\cdot, \cdot)) < s_i((\mathbf{x}^{(t)}, \mathbf{a}^{(t)}), \mathbf{y}; \tilde{g}^{(k)}(\cdot, \cdot)) \right\} \right\}$$

end for

Output: Predictions sets $\{\mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}) : i \in [m]\}$.

where $\hat{A}^*(\cdot, \cdot; \mathbf{v}) : (\mathbf{x}, \mathbf{a}) \mapsto \hat{A}(\mathbf{x}, \min(\mathbf{a}, \mathbf{v}))$ is a modified base predictor, and similarly, $\hat{R}^*(\hat{A}^*(\cdot, \cdot; \mathbf{v}), \cdot; \mathbf{v}) : (\mathbf{x}, \mathbf{a}) \mapsto \hat{R}(\hat{A}^*(\mathbf{x}, \mathbf{a}; \mathbf{v}), \min(\mathbf{a}, \mathbf{v}))$ is its reconciled version.

Second, instead of considering the modified reconciled predictor evaluated with infrastructure activity $\mathbf{v}^{(k)}$, we introduce a reconciled predictor with random infrastructure activity

$$\tilde{g}^{(k)}(\cdot, \cdot) = \mathcal{G}((\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{z}^{(k)})_{k \in \text{Cal}}, (\mathbf{x}^{(t)}, \mathbf{a}^{(t)}, \mathbf{y}^{(t)}); \boldsymbol{\nu}^{(k)}) = \hat{R}^*(\hat{A}^*(\cdot, \cdot; \boldsymbol{\nu}^{(k)}), \cdot; \boldsymbol{\nu}^{(k)}),$$

where $\boldsymbol{\nu}^{(k)}$ is drawn independently from $\{(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{z}^{(k)})_{k \in \text{Cal}}, (\mathbf{x}^{(t)}, \mathbf{a}^{(t)}, \mathbf{y}^{(t)})\}$, and $\mathcal{G}$ is a randomized algorithm¹ that outputs a pre-determined function that is independent from the input data. Hence, setting $\boldsymbol{\nu}^{(k)} = \mathbf{a}^{(k)}$ for each $k \in \text{Cal}$, which is a valid choice since $\mathbf{a}^{(k)}$ is independent from $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(t)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal} \cup \{t\}}$ by Assumption 2 and 3, we can rewrite the term involved in the third step of Algorithm 2 as

$$(19) \quad \# \left\{ k \in \text{Cal} : s_i((\mathbf{u}^{(k)}, \mathbf{a}^{(t)}), \mathbf{z}^{(k)}; \tilde{g}^{(k)}(\cdot, \cdot)) < s_i((\mathbf{x}^{(t)}, \mathbf{a}^{(t)}), \mathbf{y}; \tilde{g}^{(k)}(\cdot, \cdot)) \right\}.$$

As a consequence, this specific instance of Algorithm 3 is equivalent to Algorithm 2.

Step 3: Conditional Coverage. The same argument and notations as in the proof of Theorem A.2 conclude. Indeed, since $\mathcal{G}$ treats symmetrically all the observations within the set of observations with substituted infrastructure activity, we obtain that

$$(20) \quad \begin{aligned} M_i^*(\sigma(p), \sigma(q)) &= \mathcal{M}_i^*((\mathbf{x}^{(k)}, \mathbf{a}^{(t)}, \mathbf{y}^{(k)})_{k \in \text{Cal} \cup \{t\}}; (\mathbf{a}^{(k)})_{k \in \text{Cal}})(\sigma(p), \sigma(q)) \\ &= \mathcal{M}_i^*((\mathbf{x}^{(\sigma(k))}, \mathbf{a}^{(t)}, \mathbf{y}^{(\sigma(k))})_{k \in \text{Cal} \cup \{t\}}; (\mathbf{a}^{(\sigma(k))})_{k \in \text{Cal}})(p, q), \end{aligned}$$

where $\mathcal{M}_i^*$ the function that maps the observations with substituted infrastructure activity and the true random infrastructure activity patterns to a comparison matrix M_i^* , based on the equivalence obtained in (19), in a similar way as in the proof of Theorem A.2.

Therefore, the independence of the true infrastructure activity patterns $\{\mathbf{a}^{(k)}\}_{k \in \text{Cal}}$ from the observations with substituted infrastructure activity $\{(\mathbf{x}^{(k)}, \mathbf{a}^{(t)}, \mathbf{y}^{(k)})\}_{k \in \text{Cal} \cup \{t\}}$, the exchangeability of the latter conditional on $\mathbf{a}^{(t)}$, and the expression obtained in (20), entails that, conditional on $\mathbf{a}^{(t)}$, we have

$$(M_i^*(\sigma(p), \sigma(q)))_{p, q \in \text{Cal} \cup \{t\}} \stackrel{d}{=} \mathcal{M}_i^*((\mathbf{x}^{(k)}, \mathbf{a}^{(t)}, \mathbf{y}^{(k)})_{k \in \text{Cal} \cup \{t\}}; (\mathbf{a}^{(k)})_{k \in \text{Cal}}) = M_i^*.$$

¹The randomization requires to adapt the argument from the proof of Theorem A.2 to the case of randomized predictors, which we omit for the sake of simplicity (see Zaffran et al., 2024, Appendix C for more details).

Finally, we conclude by using the same argument as in the proof of Theorem A.2, and get the infrastructure activity-conditional guarantee in (18), as announced. $\square$

REMARK 4. The conditional coverage guarantee of Theorem A.3 holds for any reconciliation Algorithm $\mathcal{R}$, including the trivial one that does not perform any reconciliation step. Therefore, the assertion of Theorem A.3 is not that conditional coverage is guaranteed thanks to the infrastructure-aware reconciliation step, but rather that conditional coverage is a consequence of the assumptions of Theorem A.3, and is preserved even when incorporating such a reconciliation step.

B. IMPLEMENTATION OF THE ALGORITHMS

This Supplementary Material provides a detailed description of the practical implementation of the proposed algorithms. First, Section B.1 describes the construction of the base and reconciled predictors, involved in the first step of both Algorithms 1 and 2, from Algorithm 4, which is summarized in Algorithm 4. Then, Section B.2 presents the implementation of the CP-MNR-Nested* procedure. Finally, Section B.3 is devoted to the implementation of the Adaptive CP-MNR procedure.

B.1. Construction of the base and reconciled predictors. The construction of the base and reconciled predictors is summarized in Algorithm 4. The formulation encompasses the three predictors introduced in Section 3.1 and considered in the experiments of Section 5, namely the Direct (no reconciliation), OLS and rOLS predictors.

In Algorithm 4, the learning algorithm $\mathcal{A}$ (e.g., independent instances of a regression algorithm for each node) is fitted on the subset indexed by Train of the available data $\mathcal{D}$, and outputs a predictor $\hat{A} \in \mathcal{F}$, where $\mathcal{F} = \{f : \mathcal{X} \times \{0, 1\}^m \rightarrow \mathbb{R}^m\}$. Given the base predictor $\hat{A}$, the reconciliation algorithm $\mathcal{R}$ (e.g., rOLS) is fitted on the same subset of $\mathcal{D}$ indexed by Train , and outputs a reconciliation function $\hat{R} : \mathbb{R}^m \times \{0, 1\}^m \rightarrow \mathbb{R}^m$. The combination of a base predictor and a reconciliation function yields a reconciled predictor $\tilde{A} \in \mathcal{F}$, defined as $\tilde{A} : (x, a) \mapsto \hat{R}(\hat{A}(x, a), a)$.

The reconciliation algorithms considered in this article are constant, in the sense that they do not depend on the training data. The above formalism thus covers these three cases, but also allows for more general reconciliation algorithms that may depend on the training data. For instance the MinT reconciliation procedure (Wickramasuriya, Athanasopoulos and Hyndman, 2019) consists of a projection matrix that is computed from the data by computing the empirical covariance matrix of the forecast error. The explicit form of the reconciliation function $\hat{R}$ for the three considered strategies is as follows.

The Direct strategy leaves base forecasts unchanged. In this case, the reconciliation algorithm $\mathcal{R}$ is constant and outputs

$$\text{(Direct)} \quad \hat{R} : (\hat{y}, a) \mapsto \hat{y}.$$

The OLS strategy reconciles forecasts by projecting the base forecasts onto the coherent subspace $\text{Im}(H)$. In this case, let $P = H(H^\top H)^{-1}H^\top$ be the orthogonal projection onto $\text{Im}(H)$, and consider the constant reconciliation algorithm $\mathcal{R}$ that outputs

$$\text{(OLS)} \quad \hat{R} : (\hat{y}, a) \mapsto P \hat{y}.$$

The rOLS strategy reconciles forecasts by projecting the base forecasts onto the refined coherent subspace $\text{Im}(H^{(a)})$. In this case, the reconciliation algorithm $\mathcal{R}$ is constant and outputs the reconciliation function

$$\text{(rOLS)} \quad \hat{R} : (\hat{y}, a) \mapsto P^{(a)} \hat{y},$$

Algorithm 4 Construction of the predictors

Input: Structural matrix H ; dataset $\mathcal{D} = \{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)}) : k \in \text{Train} \cup \text{Cal}\}$; set of indices Train for training; learning algorithm $\mathcal{A}$; reconciliation algorithm $\mathcal{R}$.

1: Fit $\mathcal{A}$: $\hat{A}(\cdot, \cdot) \leftarrow \mathcal{A}(\{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)}) : k \in \text{Train}\})$

2: Fit $\mathcal{R}$: $\hat{R}(\cdot, \cdot) \leftarrow \mathcal{R}(\{\hat{A}(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}), \mathbf{a}^{(k)}, \mathbf{y}^{(k)} : k \in \text{Train}\}, H)$

Output: Fitted predictor $\hat{A}$, fitted reconciliation function $\hat{R}$, and fitted reconciled predictor $\tilde{A} = \hat{R}(\hat{A}(\cdot, \cdot), \cdot)$.

where $P^{(a)} = H^{(a)}((H^{(a)})^\top H^{(a)})^\dagger (H^{(a)})^\top$ with $(\cdot)^\dagger$ being the Moore–Penrose pseudo-inverse.

B.2. Implementation of CP-MNR-Nested*. First, the computation of the conformal prediction set in Algorithm 2 is made explicit from a practical standpoint. Then, several practical aspects of the implementation are discussed, including the nature of the resulting prediction set and the choice of the calibration indices.

This implementation follows the efficient implementation of cross-conformal prediction sets (Gupta, Kuchibhotla and Ramdas, 2022, Algorithm 1). Without loss of generality, the conformal prediction sets produced by CP-MNR-Nested* are of the form

$$(21) \quad \mathcal{C} = \left\{ c \in \mathbb{R} : \#\{k : r_k < |c_k - c| \} < (1 - \alpha)(N + 1) \right\},$$

for some $N \in \mathbb{N}$, some non-negative scalars $r_1, \dots, r_N$, and scalars $c_1, \dots, c_N$.

The above expression does not yield a direct implementation, but can be reformulated in terms of a family of overlapping intervals. For each $k \in [N]$, let $I_k = [c_k - r_k, c_k + r_k]$, and note that $\{r_k < |c_k - c|\} = \{c \notin I_k\}$ for all $c \in \mathbb{R}$. Thus, it follows that

$$(22) \quad \mathcal{C} = \left\{ c \in \mathbb{R} : \#\{k : c \in I_k\} > \alpha(N + 1) - 1 \right\}.$$

From (22), the computation of the conformal set reduces to identifying the regions of $\mathbb{R}$ where the function $c \mapsto \sum_{k=1}^N \mathbf{1}_{c \in I_k}$ exceeds a fixed threshold. This function is piecewise constant with changes only at the interval endpoints $\{c_k - r_k, c_k + r_k\}_{k=1}^N$. Consequently, sorting these $2N$ breakpoints and performing a single pass that updates the active interval count, thereby yield all connected components of the set $\mathcal{C}$.

REMARK 5. The computation of the prediction sets itself in CP-MNR-Nested* is not computationally intensive. On the contrary, the computation of the reconciled forecasts for each calibration point and each test point can be computationally expensive, as it requires to compute a projection matrix for each calibration point and each test point.

REMARK 6. In Algorithm 2, the practitioner can pick any subset of calibration indices $\widetilde{\text{Cal}} \subseteq \text{Cal}$ to mitigate the issue that would arise from the potential sparsity of the augmented infrastructure activity patterns.

REMARK 7. In the experiments, the subsampling is performed by selecting the calibration observations with infrastructure activity vector having at most 600 different components with respect to the test one.

REMARK 8. The CP-MNR-Nested* procedure does not necessarily yield intervals. For simplicity, the convex hull of the resulting prediction set is considered in Section 5.

Algorithm 5 Adaptive CP-MNR at level $1 - \alpha$ with step size γ

Input: Dataset $\mathcal{D} = \{(\mathbf{x}^{(k)}, \mathbf{a}^{(k)}, \mathbf{y}^{(k)}) : k \in \text{Train} \cup \text{Cal}\}$; algorithms $\mathcal{A}$ and $\mathcal{R}$; set of indices Test .

Initialization: Set $\tau = 1$; $\mathcal{D}^{(\tau)} = \mathcal{D}$; $\text{Train}^{(\tau)} = \text{Train}$; $\text{Cal}^{(\tau)} = \text{Cal}$; $\boldsymbol{\alpha}^{(\tau)} = (\alpha, \dots, \alpha)^\top \in [0, 1]^m$.

for $t \in \text{Test}$ **do**

 1: **for** $i \in [m]$ **do** run CP-MNR at level $1 - \alpha_i^{(\tau)}$ on $\mathcal{D}^{(\tau)}$ and get $\mathcal{C}_i^{(t)} = \mathcal{C}_i(\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$

 1.1: Observe $\mathbf{y}^{(t)}$ and compute $\text{error}_i^{(\tau)} = \mathbf{1}_{y_i^{(t)} \notin \mathcal{C}_i^{(t)}} \times \mathbf{1}_{a_i^{(t)} = 1}$

 1.2: Update miscoverage level $\alpha_i^{(\tau+1)} = \alpha_i^{(\tau)} + \gamma \times \text{error}_i^{(\tau)}$

end for

 2: Update dataset $\mathcal{D}^{(\tau+1)} = \mathcal{D}^{(\tau)} \cup \{(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}, \mathbf{y}^{(t)})\}$

 3: Split $\mathcal{D}^{(\tau+1)}$ into $\text{Train}^{(\tau+1)}$ and $\text{Cal}^{(\tau+1)}$

 4: Update counter $\tau = \tau + 1$.

end for

Output: Predictions sets $\{\mathcal{C}_i^{(t)} : i \in [m], t \in \text{Test}\}$.

B.3. Implementation of Adaptive-CP-MNR. This section describes Adaptive-CP-MNR, the online extension of CP-MNR introduced in Section 4.3. The procedure, summarized in Algorithm 5, is a straightforward combination of CP-MNR from Algorithm 1 with the adaptive conformal inference framework of Gibbs and Candès (2021).

In Algorithm 5, a loop is performed over the test points $t \in \text{Test}$ in chronological order and a counter $\tau \geq 1$ is used to denote the current round of the loop. The first step of the procedure consists in running CP-MNR with node-specific coverage levels that are updated sequentially. The sequence is initialized at each node by the nominal level $\alpha \in [0, 1]$. For each node $i \in [m]$, we perform CP-MNR at the current level $1 - \alpha_i^{(\tau)}$, instead of the fixed nominal level $1 - \alpha$, which yields the prediction set $\mathcal{C}_i^{(t)}$. When the true value $y_i^{(t)}$ is observed, the miscoverage level is updated with the following rule:

$$(23) \quad \forall \tau \geq 1, \quad \alpha_i^{(\tau+1)} = \alpha_i^{(\tau)} + \gamma \left(\alpha - \mathbf{1}_{y_i^{(t)} \notin \mathcal{C}_i^{(t)}} \right) \mathbf{1}_{a_i^{(t)} = 1},$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function and $\gamma > 0$ is a step size parameter. The step size γ controls the adaptation speed, and is set to $\gamma = 0.02$ in our experiments. A small value of γ yield a slow adaptation to temporal drifts, while a large value of γ may lead to oscillations outside of the $[0, 1]$ interval (Gibbs and Candès, 2021). The second, third and fourth steps of the procedure consist in updating the dataset, the training and calibration sets of indices, and the counter τ for the next round of the loop.

REMARK 9. By convention, $\alpha_i^{(\tau)} \leq 0$ yields the full space and $\alpha_i^{(\tau)} \geq 1$ the empty set.

Adaptive CP-MNR inherits the classic asymptotic coverage guarantee of adaptive conformal inference (Gibbs and Candès, 2021, Proposition 4.1). Precisely, at each node $i \in [m]$, the only difference with the original adaptive conformal inference framework is that the miscoverage level is only updated if the node is active, i.e., if $a_i^{(t)} = 1$. Therefore, the following proposition is a direct application of the classic asymptotic coverage guarantee of ACI to the Adaptive CP-MNR procedure.

PROPOSITION B.1. The Adaptive CP-MNR procedure of Algorithm 5 satisfies

$$(24) \quad \forall i \in [m], \quad \left| \frac{1}{T_i} \sum_{t \in \text{Test}} \mathbf{1}_{y_i^{(t)} \notin \mathcal{C}_i^{(t)}} \mathbf{1}_{a_i^{(t)} = 1} - \alpha \right| \leq \frac{\max\{\alpha, 1 - \alpha\} + \gamma}{\gamma T_i} \quad \text{a.s.,}$$

where $T_i = \sum_{t \in \text{Test}} \mathbf{1}_{a_i^{(t)}=1}$ denotes the number of active test points for node i . In particular, if T_i goes to infinity as $\#\text{Test}$ goes to infinity, then (24) entails that

$$(25) \quad \lim_{T_i \rightarrow \infty} \frac{1}{T_i} \sum_{t \in \text{Test}} \mathbf{1}_{y_i^{(t)} \in \mathcal{C}_i^{(t)}} \mathbf{1}_{a_i^{(t)}=1} = 1 - \alpha \quad \text{a.s.}$$

PROOF. The proof is a direct application of [Gibbs and Candès, 2021](#), Proposition 4.1.

We first get by contradiction that $\alpha_i^{(\tau)} \in [-\gamma, 1 + \gamma]$ almost surely for all $\tau \geq 1$. Then use a telescopic sum argument to obtain the bound in (24). $\square$

C. ADDITIONAL NUMERICAL RESULTS

C.1. Comparison of the reconciliation procedures in terms of efficiency. The performance of the different reconciliation procedures in terms of prediction interval length are reported in Table 4, and confirm the choice of rOLS as the reconciliation procedure. The average length of the prediction intervals is computed conditionally on the node being active for each nodes, and the win rate of each reconciliation procedure is computed with respect to the Direct baseline. Hence, a win rate significantly above 50% indicates that the reconciliation procedure yields shorter prediction intervals than the Direct baseline, advocating for its use in practice if the validity rates are similar. The results are gathered by node type to exhibit differences in performance across the hierarchy.

OLS yields, at best, a slight improvement over the Direct baseline: it shortens intervals at the national and regional levels, but it is not consistently beneficial at the station level. By contrast, rOLS is uniformly better, with a clear efficiency gain across the hierarchy; for instance with a win rate of 60%, 61.6%, and 59.4% at the national, regional, and station levels, respectively, for CP-MNR. Note that CP-MNR-Nested* is not included in the comparison, since its construction is specific to the rOLS setting (the infrastructure activity vector is not taken in consideration in Direct and OLS). At the station level, the highest efficiency gain is obtained for the CP procedure, with a win rate of 81%. This efficiency gain is actually due to the fact that CP does not achieve the nominal coverage level, as discussed next.

C.2. Comparison of the reconciliation procedures in terms of validity. Any improvement in terms of efficiency is only meaningful if it does not come at the cost of coverage. In this regard, Tables 5 and 6 reports the validity rates (as defined in Section 5.3) at the nominal level $1 - \alpha = 90\%$ and at the relaxed level $1 - 2\alpha = 80\%$, respectively. Both tables demonstrate that the rOLS reconciliation procedure does not compromise the validity of the prediction intervals. It stems from improved point forecast accuracy, both on the calibration and test sets, recovering the theoretical insights from [Principato et al. \(2026\)](#). In essence, they show that, under the assumption of i.i.d. and elliptical non-conformity scores, forecast reconciliation improves the efficiency of conformal prediction sets without compromising their validity. The only exception is the CP procedure, which does not subsample the calibration data according to node activity. Therefore, the rOLS forecast errors are set to zero in the calibration set for inactive nodes, leading to a lack of coverage conditional on the corresponding node being active in the test set.

TABLE 4

Win rate of OLS and rOLS with respect to Direct in terms of prediction interval average lengths. Results are reported as mean $\pm 1.96 \times$ standard error.

Node type	Reconciliation	CP	CP-MNR	Adaptive CP-MNR
National	OLS	50.0% $\pm$ 16.7%	50.0% $\pm$ 16.7%	80.0% $\pm$ 13.3%
	rOLS	60.0% $\pm$ 16.3%	60.0% $\pm$ 16.3%	90.0% $\pm$ 10.0%
Regional	OLS	55.9% $\pm$ 2.2%	55.9% $\pm$ 2.2%	46.9% $\pm$ 1.5%
	rOLS	61.6% $\pm$ 2.4%	61.6% $\pm$ 2.4%	47.5% $\pm$ 1.5%
Station	OLS	48.0% $\pm$ 0.4%	42.4% $\pm$ 0.3%	37.7% $\pm$ 0.3%
	rOLS	81.0% $\pm$ 0.2%	59.4% $\pm$ 0.3%	57.4% $\pm$ 0.2%

TABLE 5

Validity rates at level $1 - \alpha = 90\%$ for the different reconciliation procedures. Results are reported as mean $\pm 1.96 \times$ standard error. The best-performing reconciliation procedure is shown in bold.

Node type	Reconciliation	CP	CP-MNR	Adaptive CP-MNR
National	Direct	10.0% $\pm$ 10.0%	10.0% $\pm$ 10.0%	0.0% $\pm$ 0.0%
	OLS	30.0% $\pm$ 15.3%	30.0% $\pm$ 15.3%	0.0% $\pm$ 0.0%
	rOLS	40.0% $\pm$ 16.3%	40.0% $\pm$ 16.3%	0.0% $\pm$ 0.0%
Regional	Direct	67.2% $\pm$ 0.8%	67.2% $\pm$ 0.8%	62.8% $\pm$ 1.4%
	OLS	53.8% $\pm$ 1.7%	53.8% $\pm$ 1.7%	49.1% $\pm$ 1.6%
	rOLS	51.9% $\pm$ 1.2%	51.9% $\pm$ 1.2%	47.8% $\pm$ 0.9%
Station	Direct	55.6% $\pm$ 0.3%	56.9% $\pm$ 0.2%	61.8% $\pm$ 0.3%
	OLS	51.6% $\pm$ 0.2%	55.8% $\pm$ 0.1%	60.4% $\pm$ 0.3%
	rOLS	38.1% $\pm$ 0.4%	56.2% $\pm$ 0.4%	57.6% $\pm$ 0.6%

TABLE 6

Validity rates at the relaxed level $1 - 2\alpha = 80\%$ for the different reconciliation procedures. Results are reported as mean $\pm 1.96 \times$ standard error. The best-performing reconciliation procedure is shown in bold.

Node type	Reconciliation	CP	CP-MNR	Adaptive CP-MNR
National	Direct	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
	OLS	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
	rOLS	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
Regional	Direct	97.2% $\pm$ 0.7%	97.2% $\pm$ 0.7%	100.0% $\pm$ 0.0%
	OLS	98.8% $\pm$ 0.5%	98.8% $\pm$ 0.5%	100.0% $\pm$ 0.0%
	rOLS	98.8% $\pm$ 0.5%	98.8% $\pm$ 0.5%	100.0% $\pm$ 0.0%
Station	Direct	88.7% $\pm$ 0.4%	94.2% $\pm$ 0.3%	98.3% $\pm$ 0.1%
	OLS	84.4% $\pm$ 0.4%	94.6% $\pm$ 0.3%	98.5% $\pm$ 0.1%
	rOLS	67.7% $\pm$ 0.3%	93.6% $\pm$ 0.3%	98.0% $\pm$ 0.1%