

Mechanism-dependent antagonism of auxiliary information in substation-level load disaggregation for distribution network planning

Xuanhao Mu^{*} , Kundan Thota , Nan Liu , Jianlei Liu , Thorsten Schlachter , Veit Hagenmeyer

Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Karlsruhe, Germany

HIGHLIGHTS

- A closed-form condition explains when multiplicative correction helps or harms.
- Direction accuracy is 89.9% in RMSE against a 74.7% majority-class base rate.
- Auxiliary effects reverse with the integration mechanism and base allocation.
- Grid-to-substation mapping defines the implemented criterion coordinates.
- Absolute accuracy remains study-case- and estimand-dependent.

ARTICLE INFO

Keywords:

Spatial load disaggregation
Multiplicative post-correction
Mechanism-dependent antagonism
Graph neural network
Distribution network planning

ABSTRACT

Open-source energy system models disaggregate zonal electricity demand to substations, and modellers routinely layer auxiliary information onto that allocation, assuming more information cannot hurt. The pipeline gives no warning when it does. We fix the Voronoi assignment and test 22 configurations crossing uniform, Grid Point Model (GPM), and graph neural network (GNN) weighting with post-correction, prior-loss, and feature-fusion integration on 1891 British, 143 Australian, and 13 German substations, with Germany descriptive. At unit correction intensity ($\beta = 1$), Nighttime Light and Proximity cut RMSE by 37–43% on the tested uniform and GPM bases, raise the British GNN value by 21%, and lower the Australian one by 30%. Absolute accuracy is study-case- and estimand-dependent: pooled predictive R^2 is positive for selected British configurations, whereas region-mean predictive R^2 is negative for all 22 configurations and for the pipeline-matched grid reference in both main study cases. Under geometric re-normalization and log-MSE, correction helps exactly when factor–residual alignment exceeds half the factor's relative dispersion. In the implemented pipeline it classifies original-scale RMSE direction in 391/435 configurations (89.9%, against a 74.7% majority-class base rate), rising to 97.0% outside the displacement band and 67.3% inside. We derive the mapping from grid-point correction to the criterion's substation factor. The condition characterizes error direction, not magnitude, and evaluating factor–residual alignment and residual dispersion requires metered substation demand; it informs integration-mechanism trade-offs, not absolute accuracy or downstream planning decisions. Code is available at <https://github.com/KIT-IAI/AllocateGNN>

1. Introduction

Distribution network planning requires substation-level demand estimates, although electricity consumption is often available only by zone or region. Open-source models bridge this gap with a Voronoi pipeline comprising spatial partitioning, demand weighting, and auxiliary correction [1–3]. More accurate demand inputs remain distinct from the reliability criteria and decision rules required for planning decisions.

Multiplicative post-correction scales grid-point demand by auxiliary factors and re-normalizes regional totals [4–6]. This design assumes that the uncorrected base allocation contains little spatial structure. Replacing a rule-based weighting method with a learned one may violate that assumption because the reused correction can overwrite information already encoded by the base.

Voronoi partitioning is typically fixed [7], while weighting uses uniform allocation, the Grid Point Model (GPM), or a graph neural

^{*} Corresponding author.

Email address: xuanhao.mu@kit.edu (X. Mu).

network (GNN). Auxiliary information such as Nighttime Light (NTL) intensity [8] and the substation-proximity signal (Proximity) [9] are then attached through the chosen integration mechanism. Reviews and single-axis comparisons document diverse implementations but do not test whether one correction remains appropriate after the weighting method changes [2,3,10–12]. Related modeling research likewise shows that the same domain knowledge can behave differently when encoded through different mechanisms [13–15].

This study therefore asks when the same auxiliary information helps or harms after the weighting changes from rule-based to learned, and when mechanism-dependent antagonism emerges. It varies the weighting method and the integration mechanism under a fixed Voronoi assignment, treating each configuration as an evaluation design cell rather than a ranked candidate.

The study makes three contributions:

1. The first contribution is an analytical condition for when a multiplicative auxiliary correction reduces or increases error. It recasts classical control-variate and ratio-adjustment conditions for multiplicative spatial correction [16,17]. Under geometric re-normalization and log-MSE, factor-residual alignment must exceed half the factor's relative dispersion, independently of how the base allocation was produced (Section 3.5). In the implemented pipeline at the stated numerical floor, this criterion classifies original-scale RMSE direction in 391/435 configurations (89.9% versus a 74.7% majority-class base rate), rising to 97.0% outside the displacement band (Section 5.5). Computing factor-residual alignment and residual dispersion requires metered substation demand in the evaluated region. Section 5.9 examines estimation from metered subsamples and does not establish a consistent advantage over simple decision rules.
2. The second contribution shows that the same auxiliary data can reduce or increase error depending on the integration mechanism and the base allocation. A controlled evaluation of configuration-dependent reversals uses metered demand from 1891 British and 143 Australian substations, with a descriptive 13-substation German case. At unit multiplicative intensity, the same signals reduce static-base RMSE by 37–43%, but change learned-base RMSE by +21% in Britain and –30% in Australia. Absolute predictive accuracy is reported separately and remains negative for every Australian configuration.
3. The third contribution defines the criterion's correction-factor inputs in the implemented pipeline rather than only in the idealized form of the proposition. It maps grid-point correction through arithmetic conservation and Voronoi aggregation to the effective substation factor used by the criterion (Section 3.6). The mapping matches stored factors to machine precision and separates the exact geometric/log-MSE result from the empirical arithmetic/RMSE classifier.

The contribution is not a stronger Graph Neural Network (GNN) but rather the identification of a mechanism-dependent failure mode within a reused upstream module for open-source energy system modeling and its formalization as a testable criterion. Together, the contributions provide a strict condition, configuration-level evidence about the direction, and an engineering map of the implemented pipeline. They do not establish an absolutely adequate configuration or map the disaggregation error directly to a planning margin, loading threshold, or investment action. The remainder of this paper is organized as follows: Section 2 reviews related work. Sections 3 and Section 4 describe the methodology and the experimental design. Section 5 reports results. Section 6 discusses mechanisms, practical implications, and limitations. Section 7 states conclusions.

2. Related work

Spatial disaggregation allocates regional aggregate quantities to finer spatial units. The resolution at which these units are defined affects generation siting, transmission investment, and total system cost in energy optimization models [18–21]. The present paper does not address which resolution to adopt, but rather how to disaggregate aggregate data for fine-grained nodes once a resolution has been chosen.

Raventós et al. [3] compared the disaggregation procedures of eight German transmission network models along three dimensions: spatial partitioning, demand weighting, and auxiliary information integration. There was wide divergence across all three dimensions, but the comparison remained inter-model because substation-level metered data were unavailable as ground truth. Patil et al. [10] extended this taxonomy to climate action planning, sorting approaches into proxy-variable-based decomposition, machine learning, and hybrid methods, and provided cross-country empirical evidence by applying several of these methods to end-use sector consumption in Germany and Spain [22]. However, most of the methods mentioned in both papers lack validation using ground-truth data.

Along the demand-weighting dimension, early work merely relied on deterministic allocation rules. Mei et al. [23] applied socio-demographic features to spatial estimation of electricity consumption using statistical learning. Subsequent studies distributed electricity consumption and generation capacity across voltage levels using fixed proxy variables such as population, employment, and industrial output as weights [2,12,24]. In prior work [11], we showed that a Heterogeneous Graph Neural Network (HGNN) can learn spatial weight functions from multimodal features, replacing rule-based demand weighting.

Along the spatial-partitioning dimension, Voronoi Diagram (VD) remains the default in both open-source frameworks [1] and commercial tools; this paper fixes the partition to Voronoi and varies the remaining two axes. Decomposition is inherently underdetermined: infinitely many allocations can reproduce a regional sum, which is why auxiliary information is needed to constrain the solution. Adjacent fields supply direct evidence that the integration mechanism of domain knowledge is consequential, and that different mechanisms produce distinct effects. Raventós et al. [3] identified post-correction as a procedural dimension but treated all corrections as a single category. They neither distinguish inference-time correction from training-time guidance, nor compare their effects on disaggregation accuracy. Neither static nor learning-based disaggregation methods have examined whether the same auxiliary data yields different results when it is used as an external overlay, an inference-time correction, or a training-time prior.

In physics-informed neural networks, Chen et al. [13] showed that encoding physical laws as loss-function penalties (soft constraints) enables the network to trade off physical consistency against data fit. The trade-off is governed by the loss coefficient: a low weight allows large violations, a high weight enforces near-exact compliance. Encoding the same laws architecturally (hard constraints) operates through a different mechanism: it narrows the feasible solution space before optimization begins, enforcing physical consistency by construction rather than by gradient pressure. Both mechanisms lead to distinct convergence behavior and generalization profiles, even when the underlying physical information is identical.

The same conclusion holds when domain priors are incorporated into graph-based and multi-task architectures. Liang et al. [14] compared three entry points for domain priors in knowledge-guided graph models: input features (freely weighted by the model), graph structure (constraining which node interactions are representable), and auxiliary loss terms (creating competing gradient directions that reshape learned representations). No single mechanism dominated across tasks. Gong et al. [15] reported an analogous finding in multi-task learning: hard parameter sharing forces shared low-level representations, while soft parameter

sharing permits task-specific feature spaces connected by regularization; the choice between them, together with the loss weighting strategy, shifts the balance between positive transfer and negative interference.

Taken together, the same information, entering a learning pipeline through different mechanisms, produces different model behavior. None of these results, however, have been tested in spatial disaggregation.

Across the spatial disaggregation literature, the functional form through which auxiliary information enters the allocation is overwhelmingly multiplicative. In weighted areal interpolation and dasymetric mapping, i.e., methods that redistribute zonal totals using auxiliary spatial data such as land cover, target-unit values are obtained by multiplying a base allocation by spatially varying weights derived from auxiliary data such as land use, building density, NTL, and population grids, then re-normalizing to preserve the source-zone total [4–6]. The same structure persists in hybrid pipelines that combine machine-learning density estimates with demand-conserving rescaling: the learned density field enters as a multiplicative weight, and a global scaling factor enforces conservation [25,26]. Even in emission inventory disaggregation, proxy intensities are combined multiplicatively before proportional allocation [10]. The multiplicative form owes its prevalence to a structural convenience: it guarantees non-negativity and re-normalization, and automatically satisfies the demand conservation constraint. Whether this convenience carries hidden costs when the base allocation is itself a learned model, rather than a uniform or rule-based prior, has not been examined.

NTL imagery is among the most common proxy variables in spatial disaggregation. Chen et al. [8] constructed a global 1 km × 1 km gridded dataset of gross domestic product and electricity consumption from calibrated NTL data and reported a strong positive correlation between NTL intensity and electricity consumption at the global scale; Bhattarai et al. [27] corroborated this at national and sub-national levels. NTL has also been used to disaggregate greenhouse gas emissions: Crippa et al. [28] employed it as a spatial allocation proxy in the EDGAR v8.0 database, and Danylo et al. [29] combined building footprints, population density, and energy consumption data to produce high-resolution emission distributions for the residential sector. The NTL-electricity correlation is scale-dependent. Meta-analyses report that predictive strength degrades at sub-national and sub-annual scales, where population, land use, and industrial structure become confounders [8,27].

Across these studies, NTL is included in the allocation solely as a multiplicative correction factor applied to an initial estimate. This form is natural when the initial estimate is rule-based or uniform, but a learning-based allocation model opens a second route: directly encoding the NTL signal into the training objective. Whether the two routes produce equivalent disaggregation accuracy has not been tested on ground-truth substation demand.

Substation proximity is a second auxiliary signal, motivated by the observation that spatially closer substations tend to carry more correlated loads [9,30]. Proximity overlaps with NTL: urban centers exhibit both high NTL intensity and high substation density. Whether combining both signals yields complementary gains or introduces redundant spatial correlation has not been tested. Existing work uses geographic proximity or load density as features; to our knowledge, no study has performed ablation or information-theoretic analysis to disentangle the marginal contribution of each.

The preceding review converges on one central gap. Across spatial disaggregation, auxiliary information enters the allocation almost exclusively through one mechanism (multiplicative correction of a base estimate), while the physics-informed, knowledge-guided, and multi-task learning literatures show that the entry point of domain knowledge into a learning pipeline is consequential [13–15]. Yet this has never been tested in spatial load disaggregation: the integration practices cataloged by Raventós et al. [3] have not been compared on ground-truth data, and the conditions under which the dominant multiplicative form helps or harms a learned base allocation have not been stated. The research question therefore reduces to one: does the mechanism through which the

same auxiliary information enters the pipeline change disaggregation accuracy, and if so, under what conditions? One subordinate question accompanies it. The partial overlap between NTL intensity and Proximity has been noted but not quantified; whether their joint use is synergistic or antagonistic may itself depend on the integration mechanism, which makes this signal pair a test instance of the central question rather than an independent gap.

The paper addresses this hierarchy directly: the controlled comparison of integration mechanisms and signal combinations, validated against substation-level metered demand, answers the central question; the conditional-benefit criterion (Section 3.5, tested in Section 5.5) states the conditions; and the NTL–Proximity pair carries the subordinate question through every experiment.

3. Methodology

The study uses a controlled two-axis evaluation: Voronoi assignment is fixed while demand weighting and the auxiliary-information integration mechanism are crossed. Each reported configuration is a design cell, not an entry in a method-ranking exercise.

3.1. Problem definition

Spatial load disaggregation distributes aggregate regional electricity demand to individual primary substations. The inputs consist of M administrative regions $\mathcal{R} = \{r_1, \dots, r_M\}$ with known aggregate demand D_r ; N primary substations $\mathcal{T} = \{t_1, \dots, t_N\}$ with geographic coordinates (x_j, y_j) and metered peak demand d_j^{actual} ; and a regular grid of P spatial sample points, hereafter *grid points*, $\mathcal{A} = \{a_1, \dots, a_P\}$ covering the study area, each carrying a multi-source feature vector $\mathbf{f}_a$. Throughout, *grid* denotes this spatial sampling lattice, whereas an electrical system is called a *network*.

The objective is to recover a two-layer allocation mapping:

$$\hat{d}_j = \sum_{r \in \mathcal{R}} D_r \cdot \pi_{r \rightarrow j}, \quad \forall j \in \mathcal{T} \quad (1)$$

where $\pi_{r \rightarrow j}$ denotes the fraction of region r 's demand routed to substation j , subject to $\sum_j \pi_{r \rightarrow j} = 1$ and $\pi_{r \rightarrow j} \geq 0$ for all r .

An intermediate grid-point layer decomposes this transfer into two steps:

$$\pi_{r \rightarrow j} = \sum_{a \in \mathcal{A}_r} w_{ra} \cdot w_{aj} \quad (2)$$

where w_{ra} distributes demand from region r to grid point a , w_{aj} routes demand from grid point a to substation j , and $\mathcal{A}_r$ is the set of grid points within region r .

Four index sets recur throughout and determine the domain of every normalization, loss, and conservation constraint; we fix them here. Each region r is partitioned into administrative sub-regions (International Territorial Level (ITL)-3 units in the British case), collected in S_r ; the sub-regions are the units at which aggregate demand is known, with totals D_s satisfying $D_r = \sum_{s \in S_r} D_s$. The grid points inside region r form $\mathcal{A}_r \subseteq \mathcal{A}$, and those inside sub-region s form $\mathcal{A}_s$, so that $\mathcal{A}_r = \bigcup_{s \in S_r} \mathcal{A}_s$ with the union disjoint. In the GNN graph, each sub-region s is represented by one sub-region node, and the graph is a star through administrative containment: the neighborhood of sub-region node s is exactly its grid points, $\mathcal{N}(s) = \mathcal{A}_s$. The set $\text{RCI} \subseteq \mathcal{A}$ contains grid points whose residential, commercial, and industrial land-use shares sum to more than 0.5; it is the support on which correction-factor statistics and prior targets are computed. Finally, $\mathcal{T}_r \subseteq \mathcal{T}$ restricts Voronoi candidates to substations inside region r . With these sets, the scopes are: the softmax in Eq. (6) normalizes over $\mathcal{N}(s) = \mathcal{A}_s$; the losses in Eqs. (7) and (22) average over sub-region nodes, the prior target being supported on $\mathcal{A}_s \cap \text{RCI}$; the post-correction in Eq. (11) re-normalizes within each $\mathcal{A}_s$ and conserves D_s , hence also D_r ; and the Voronoi assignment in Eq. (4) searches over $\mathcal{T}_r$.

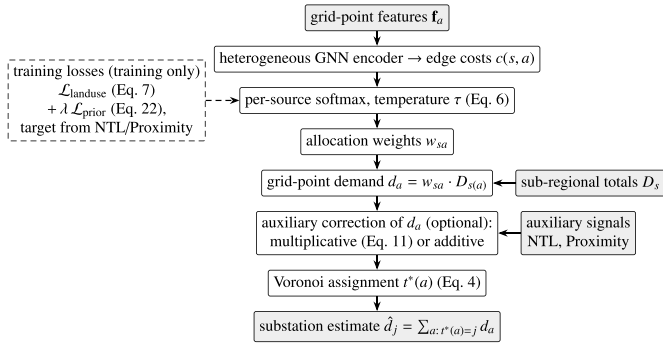


Fig. 1. End-to-end computation of the GNN pipeline. Solid arrows form the inference path; dashed elements act only during training. The prior-loss mechanism enters through the training objective (left), whereas post-correction rescales grid-point demands after inference and before Voronoi aggregation.

All evaluated configurations are expressed as a combination of three orthogonal design axes:

$$\text{Configuration} = \underbrace{\text{Spatial partitioning}}_{\text{Voronoi}} \times \underbrace{\text{Demand weighting}}_{\text{Uniform / GPM / GNN}} \times \underbrace{\text{Auxiliary information integration}}_{\text{None / NTL / Prox / NTL+Prox}} \quad (3)$$

Eq. (3) assumes independence among the three axes; the experiments in Section 5.3 test whether this holds.

For the third axis, auxiliary information enters the allocation through one of three mechanisms:

1. **Grid Point Model (GPM)-based post-correction:** multiplicative factors applied after GPM demand shares are computed, adjusting the allocation before Voronoi aggregation.
2. **GNN-based post-correction:** the same multiplicative factors, but applied to GNN outputs after inference.
3. **GNN-based prior loss:** additional loss terms injected during GNN training that encode auxiliary spatial patterns, with no modification at inference time.

Fig. 1 traces the end-to-end computation for the GNN pipeline, from raw inputs to the substation-level estimate, and marks the exact points at which the two integration mechanisms enter: the prior loss acts on the training objective, whereas post-correction rescales demands after inference and before aggregation. Every symbol in the figure is defined in the subsection where the corresponding stage is described.

3.2. Spatial partitioning

Spatial partitioning determines which substation each grid point feeds into. Throughout this paper, partitioning follows Voronoi nearest-neighbor assignment, the default method in both open-source and commercial power system simulation tools [1,3].

Each grid point a is assigned to the geographically closest substation in projected Euclidean distance:

$$r^*(a) = \arg \min_{j \in \mathcal{T}_r} \|c_a - c_j\|_2 \quad (4)$$

where c_a and c_j are two-dimensional projected coordinates and $\mathcal{T}_r$ restricts candidates to substations within region r . This produces a hard assignment: $w_{aj} = 1$ if $j = r^*(a)$ and zero otherwise. The predicted demand at substation j then equals the sum of grid-point demands within its Voronoi cell:

$$\hat{d}_j = \sum_{a: r^*(a)=j} d_a \quad (5)$$

This aggregation step lies outside every training objective considered in this paper: the GNN is trained to allocate sub-region demand to grid points (Section 3.3), whereas evaluation compares $\hat{d}_j$ with metered substation demand after the hard assignment in Eq. (4). A reallocation of weights that leaves the training loss nearly unchanged can therefore still move grid-point demand across Voronoi cell boundaries and change substation-level error; the magnitude of this freedom is measured by the isoline analysis in Section 5.5.

3.3. Demand weighting

Demand weighting determines the share of regional demand each grid point carries. Three weighting methods are compared: uniform weighting as a reference, and two published approaches, the Grid Point Model (GPM) [31] and an HGNN [11].

Uniform weighting splits regional demand equally among all grid points in the region: $d_a = D_r / |\mathcal{A}_r|$.

The GPM weighting method [31] assigns each grid point to its dominant land-use class $k^*(a) = \arg \max_k p_a^{(k)}$ and weights it by the regional consumption share of that class, $\text{pct}_r^{(k^*)}$. Weights are normalized per region to conserve D_r . Residentially dominated grid points, therefore, experience higher demand in regions where residential consumption accounts for a large fraction, and, conversely, other classes experience lower demand.

The GNN weighting method [11] frames allocation as edge-weight prediction on a heterogeneous graph whose node types are sub-region nodes and grid-point nodes. A Heterogeneous Graph Transformer encoder [32] produces node embeddings; allocation weights follow from a temperature-scaled softmax over learned edge costs, grouped by source:

$$w_{sa} = \frac{\exp(-c(s, a)/\tau)}{\sum_{a' \in \mathcal{N}^+(s)} \exp(-c(s, a')/\tau)}, \quad d_a^{\text{GNN}} = w_{s(a), a} \cdot D_{s(a)} \quad (6)$$

where $c(s, a)$ is the learned cost of the edge connecting sub-region node s to grid point a , τ is a learnable temperature parameter (with parameterization and initialization in Appendix A.3), $\mathcal{N}^+(s) = \mathcal{A}_s$ is the neighborhood of sub-region node s (Section 3.1), $s(a)$ denotes the sub-region containing grid point a , and $D_{s(a)}$ is that sub-region's aggregate demand. Because the softmax normalizes within $\mathcal{A}_s$, the resulting demands satisfy $\sum_{a \in \mathcal{A}_s} d_a^{\text{GNN}} = D_s$ by construction.

Training is self-supervised via a land-use reconstruction loss:

$$\mathcal{L}_{\text{landuse}} = \frac{1}{|\mathcal{V}_s|} \sum_s D_{\text{KL}}(\mathbf{q}_s^{\text{true}} \parallel \text{norm}(\sum_{a \in \mathcal{N}^+(s)} w_{sa} \cdot \mathbf{M}_{a,:})) \quad (7)$$

where $\mathcal{V}_s$ is the set of sub-region nodes, $\mathbf{q}_s^{\text{true}}$ is the observed consumption-share vector of sub-region s derived from regional economic statistics (detailed in Section 4), and $\mathbf{M}_{a,:}$ is the one-hot indicator of grid point a 's dominant land-use class $k^*(a)$, assigned during graph construction as in Section 3.3. The loss penalizes deviations between the observed regional consumption structure and the structure reconstructed from weighted grid-point land-use vectors. Because each grid point contributes its weight to exactly one class, the reconstructed mixture aggregates the weight vector into at most five class masses per sub-region (fewer where a class is absent), so Eq. (7) constrains at most five linear functionals of the weights and is invariant to any reallocation that preserves these class masses. Probability floors and zero handling in the implementation are collected in Appendix A.5.

These aggregate constraints leave the weights under-determined: any reallocation among grid points sharing a dominant class preserves every class mass and hence the value of Eq. (7), so the loss alone does not identify w_{sa} as a spatial demand share; in strict terms, w_{sa} is a proxy-supervised allocation weight. Two structural properties narrow the solution space that the trained GNN can actually reach. First, the parameterization: weights are produced by one encoder and edge-cost head shared across all sub-regions and regions (Eq. 6; Appendix A), so

w_{sa} is a deterministic function of grid point a 's land-use profile within its sub-region. Grid points with identical profiles therefore receive identical weights (loss-invisible reallocations between them are not expressible at all), and reallocations across differing profiles must be realized by a globally shared function rather than tuned freely per sub-region. Second, in prior-loss configurations, the auxiliary target of Eq. (21) adds independent constraints on the RCI support. The isoline analysis in Section 5.5 samples allocations on exact isolines of Eq. (7) to place the learned solution within one finite Monte Carlo reference distribution. Because the sampler neither exhausts the isoline nor establishes reachability under the shared parameterization, it does not bound the full error freedom left by the loss. We accordingly interpret w_{sa} throughout as a learned allocation weight whose demand semantics are established empirically rather than assumed from the training objective.

Two consequences of this self-supervised design are relevant to the integration analysis. First, feature-level fusion (concatenating NTL or Proximity values onto the five-dimensional node feature vector) acquires no dedicated learning channel: Eq. (7) measures only land-use reconstruction quality, so no gradient signal rewards NTL-informed allocation and the GNN can discard features that do not reduce the reconstruction error; what the injected features actually do to the allocation is therefore an empirical question, which the fusion configurations in the results section measure in both main study cases. Second, the prior-loss mechanism (Section 3.8) is the only one that writes auxiliary information into the training objective, generating a gradient signal aligned with the auxiliary pattern. Post-correction (Section 3.5) operates outside the training loop entirely and is therefore decoupled from the GNN's learned representation.

These consequences presuppose a precise account of which spatial or geometric information reaches the GNN at all; the account below is generated programmatically from the training code and the cached graphs. Grid-point node features consist of the five land-use proportions and nothing else (no coordinates and no NTL radiance); the graphs contain zero direct edges between grid points, connectivity being limited to the membership edges that link each grid point to its sub-region source; and coordinates, NTL, Proximity, the RCI mask never enter the encoder's forward pass. The GNN thus receives neither locations nor distances through any input channel, so any proximity information in its representation would have to be inferred from land-use composition alone, a premise tested directly in Section 5.4. Table A.14 itemizes all eleven channels.

3.4. Auxiliary information

Two auxiliary information sources are examined: NTL intensity and proximity. Both are entered as multiplicative correction factors.

The NTL correction factor uses radiance as a spatial proxy for human activity [8,27]. The correction factor takes the form:

$$\text{ntl_factor}(a) = \frac{\log(1 + \text{ntl}(a) + \epsilon)}{\log(1 + \tilde{m}_{\text{ntl}})} \quad (8)$$

where $\text{ntl}(a)$ is the radiance at grid point a (nanowatts/cm²/sr), ϵ is the 5th percentile of non-zero NTL values among Residential + Commercial + Industrial (RCI) grid points, and $\tilde{m}_{\text{ntl}}$ is the median NTL over RCI grid points.

The floor ϵ prevents dark areas (e.g., agricultural or forested land) from collapsing to zero while leaving the relative ordering among lit grid points intact. The logarithmic transform compresses the wide dynamic range of raw radiance, so the factor is stable under moderate perturbations of ϵ .

The Proximity correction factor uses the tendency of primary substations to be sited near load centers, a premise embedded in siting practice itself, where placement minimizes the distance between substations and the loads they serve [33,34]. A grid point surrounded by many substations is therefore more likely to lie in a high-demand zone. The Proximity

score sums inverse-distance contributions from all substations:

$$\text{prox}(a) = \sum_{j \in \mathcal{T}} d(a, j)^{-\gamma} \quad (9)$$

where $d(a, j) = \max(\|\mathbf{c}_a - \mathbf{c}_j\|_2, 0.01 \text{ km})$ is the projected Euclidean distance clamped at 10 m to prevent numerical overflow.

With $\gamma = 2$, contributions decay fast enough that only nearby substations carry appreciable weight. The correction factor follows the same log-ratio form as the NTL factor:

$$\text{prox_factor}(a) = \frac{\log(1 + \text{prox}(a))}{\log(1 + \tilde{m}_{\text{prox}})} \quad (10)$$

with $\tilde{m}_{\text{prox}}$ being the median Proximity score over RCI grid points.

3.5. Multiplicative post-correction

Both the GPM and GNN pipelines can incorporate auxiliary information through the same mechanism: grid-point demands are multiplied by a correction factor and re-normalized within each sub-region s to conserve D_s (and therefore D_r). Given a base demand d_a^{base} and a correction factor $f(a)$, the corrected demand is:

$$d_a^{\text{corr}} = D_{s(a)} \cdot \frac{d_a^{\text{base}} \cdot f(a)}{\sum_{a' \in \mathcal{A}_{s(a)}} d_{a'}^{\text{base}} \cdot f(a')} \quad (11)$$

where $f(a)$ is $\text{ntl_factor}(a)$, $\text{prox_factor}(a)$, or both applied sequentially, and the normalization domain $\mathcal{A}_{s(a)}$ is the sub-region of a as defined in Section 3.1.

The multiplicative form of Eq. (11) is the prevailing choice in the disaggregation literature (Section 2); an additive form is defined in Section 3.7 and evaluated alongside it.

For GNN-based configurations, the base allocation w_{sa} is already normalized by the per-source softmax (Eq. 6). Eq. (11) therefore introduces a second normalization step. This second normalization guarantees demand conservation. It does, however, depart from the learned allocation, and this departure is caused by the factor $f(a)$ itself, which rescales the inter-point demand ratios encoded by the softmax in proportion to its own dispersion; the re-normalization is provably neutral with respect to allocation shape. High-variance correction factors (e.g., NTL in regions with mixed urban and rural land) produce larger departures than low-variance ones.

Proposition 1 (Shape neutrality of re-normalization). Fix a sub-region s and let $d_a^{\text{base}} > 0$ and $f(a) > 0$ for $a \in \mathcal{A}_s$. The corrected demands of Eq. (11) satisfy

$$d_a^{\text{corr}} = \left(D_s / \sum_{a'} d_{a'}^{\text{base}} f(a') \right) \cdot d_a^{\text{base}} f(a), \quad (12)$$

i.e., Eq. (11) acts on each sub-region as multiplication by a single positive constant. Consequently, (i) all demand ratios within $\mathcal{A}_s$ equal those of the unnormalized products $d_a^{\text{base}} f(a)$, and (ii) a variant that skips the re-normalization differs from Eq. (11) exactly by these sub-regional constants.

The proposition (proof in Appendix E) resolves the concern that a "no re-normalization" diagnostic confounds normalization with conservation: the two variants differ only by conservation offsets, so the second normalization itself cannot distort the learned allocation. The identity holds exactly and is confirmed to machine precision on all pipeline outputs.

Whether this factor-induced departure helps or harms is not a structural property of the multiplicative form; it depends on how well the correction factor tracks the residuals of the base estimate. The following criterion makes the condition precise (proof in Appendix E).

Proposition 2 (Conditional-benefit criterion). In its strict domain, let n evaluation units carry positive true values y_i and positive base estimates b_i ,

and let $f_i > 0$ be a correction factor. Write $r_i = \log(y_i/b_i)$ for the log residual ratios and let c_i denote the centered $\log f_i$. If the corrected estimate multiplies b_i by f_i and re-normalizes geometrically, then, and only in this geometric/log-loss setting, the mean squared error in log space strictly decreases if and only if

$$2\langle \mathbf{r}, \mathbf{c} \rangle > \|\mathbf{c}\|^2, \quad \text{equivalently} \quad \rho > \frac{\sigma_c}{2\sigma_{\text{res}}}, \quad (13)$$

where ρ is the Pearson correlation of (c_i, r_i) and $\sigma_c, \sigma_{\text{res}}$ are their population standard deviations (the subscripts denote correction and residual, respectively).

In words, $\mathbf{r}$ collects what the base gets wrong (its log residuals against the truth), $\mathbf{c}$ collects what the correction proposes to change (its centered log factor), and ρ measures how well the proposed change tracks the actual error.

Four remarks position the criterion. First, it is the disaggregation analog of Cochran's classical condition for ratio estimators, $\rho > C_x/(2C_y)$ [16]: a multiplicative correction is a ratio adjustment, and it reduces error only when the factor is sufficiently correlated with what it is meant to fix, relative to its own dispersion. Second, the factor $1/2$ in the threshold has a mechanical origin: the correction replaces the log-space error $\|\mathbf{r}\|^2$ by $\|\mathbf{r} - \mathbf{c}\|^2 = \|\mathbf{r}\|^2 - 2\langle \mathbf{r}, \mathbf{c} \rangle + \|\mathbf{c}\|^2$, so it earns the alignment term twice but pays its own squared magnitude once. Rewriting $2\langle \mathbf{r}, \mathbf{c} \rangle > \|\mathbf{c}\|^2$ through ρ, σ_c , and σ_{res} places the 2 under the threshold (Appendix E.2). Third, Proposition 2 is exact for geometric re-normalization and log-MSE, whereas the implemented pipeline uses arithmetic mass conservation, Voronoi aggregation, and RMSE on the original demand scale. These changes move the empirical boundary; Section 5.5 therefore evaluates the threshold as a fixed-rule predictor under a stated numerical-floor protocol rather than extending the if-and-only-if statement to the implemented loss. Fourth, the inequality itself is not new: with the gain fixed at one, it is the classical condition for a control variate to reduce variance [17], and the first remark places it beside Cochran's ratio-estimator bound. What the paper adds is the correction-operator formulation and its domain of validity: the correction acts multiplicatively under a mass-conservation constraint (neutral by Proposition 1). Unit gain is a registered design choice rather than a consequence of label availability; label-free, data-dependent intensity mappings are possible, as the additive scale below demonstrates. Section 5.6 therefore reports the multiplicative intensity sweep separately, while Sections 5.5 and Section 5.9 estimate and test the condition's inputs across bases and case studies. The strict statement presupposes positive true values, base estimates, and factors and measures error in log space. It is agnostic to how the base estimates were produced, but its application to the arithmetic, aggregated, linear-RMSE pipeline is an empirical classification problem whose accuracy is reported separately. Eq. (13) defines the strict log-domain boundary; Section 5.5 measures how accurately that boundary classifies correction direction in the implemented pipeline.

Within the strict grid-unit/geometric/log-MSE domain, a corollary for the combined correction follows by substitution. Applying the NTL and Proximity factors in sequence is, by Proposition 1, a single multiplicative correction by the product $f_N(a)f_P(a)$ up to sub-regional constants; the centered log factors therefore add, $\mathbf{c} = \mathbf{c}_N + \mathbf{c}_P$, and the strict condition becomes

$$2\langle \mathbf{r}, \mathbf{c}_N \rangle + 2\langle \mathbf{r}, \mathbf{c}_P \rangle > \|\mathbf{c}_N\|^2 + \|\mathbf{c}_P\|^2 + 2\langle \mathbf{c}_N, \mathbf{c}_P \rangle. \quad (14)$$

The cost side contains the grid-unit log-domain cross-term $2\langle \mathbf{c}_N, \mathbf{c}_P \rangle$. This identity does not commute with Voronoi aggregation: the effective combined substation factor $F_{NP,j}$ is generally not the product of factors obtained by aggregating N and P separately. Eq. (14) is therefore not used to predict or explain post-aggregation original-scale RMSE. Section 5.3 reports the positive grid-factor association and the observed RMSE interaction as separate empirical facts.

3.6. From grid-point correction to substation-level factors

The implemented correction acts at grid points, whereas the criterion is evaluated where ground truth is available, at substations. Let grid point a belong to the normalization sub-region $s(a)$: with base demand b_a , positive auxiliary factor f_a , and known sub-regional total D_s . On every positive-mass domain, the arithmetic conservation constant and corrected grid demand are

$$C_s = \frac{D_s}{\sum_{a \in \mathcal{A}_s} b_a f_a}, \quad d_a^{\text{corr}} = C_{s(a)} b_a f_a. \quad (15)$$

For a zero-demand domain ($D_s = 0$), the implementation sets all corrected grid demands to zero. The German case contains 23 such domains; no positive-demand domain has a zero denominator. At the substation level used by the criterion, all 13 German truths, base predictions, and corrected predictions are positive, so none of the 15 German configuration rows is epsilon-protected.

Let V_j be the set of grid points assigned to substation j by the Voronoi partition. The base and corrected substation predictions are

$$b_j = \sum_{a \in V_j} b_a, \quad b'_j = \sum_{a \in V_j} C_{s(a)} b_a f_a. \quad (16)$$

The effective substation-level correction factor used in Figs. 4 and 5, and 7 is therefore

$$F_j = \frac{b'_j + \varepsilon}{b_j + \varepsilon} = \frac{\sum_{a \in V_j} C_{s(a)} b_a f_a + \varepsilon}{\sum_{a \in V_j} b_a + \varepsilon}, \quad (17)$$

where $\varepsilon = 10^{-6} \sum_j y_j$ is the same region-specific numerical safeguard used by the analysis code. With

$$R_j = \frac{y_j + \varepsilon}{b_j + \varepsilon}, \quad r_j = \log R_j, \quad \ell_j = \log F_j, \quad c_j = \ell_j - \bar{\ell}, \quad (18)$$

the empirical coordinates are $\rho = \text{Corr}(\ell, r)$, $\sigma_c = \text{Std}(\ell)$, and $\sigma_{\text{res}} = \text{Std}(r)$. Correlation and standard deviation are unchanged by centering ℓ . The implemented arithmetic correction changes the log residual from r_j to $r_j - \ell_j$, however, not automatically to $r_j - c_j$; this is the precise separation between the strict geometric proposition and the empirical pipeline predictor.

3.7. Additive post-correction

The additive form perturbs the base allocation by a centered, rescaled version of the same correction factor instead of multiplying by it. For grid point a in sub-region $s = s(a)$:

$$d_a^{\text{add}} = D_s \cdot \frac{[d_a^{\text{base}} + \alpha_s (f(a) - \bar{f}_s)]_+}{\sum_{a' \in \mathcal{A}_s} [d_{a'}^{\text{base}} + \alpha_s (f(a') - \bar{f}_s)]_+} \quad (19)$$

where $\bar{f}_s$ is the mean of f over $\mathcal{A}_s$, $[x]_+ = \max(x, 10^{-6})$ enforces non-negativity, and the scale

$$\alpha_s = \frac{\text{std}_{a \in \mathcal{A}_s}(d_a^{\text{base}})}{\text{std}_{a \in \mathcal{A}_s}(f(a) - \bar{f}_s)} \quad (20)$$

matches the dispersion of the additive term to that of the base allocation (population standard deviations; $\alpha_s = 0$ if either dispersion vanishes).

This definition fixes the five properties required of a formal correction relation. First, α_s is a label-free, sub-region-specific intensity adaptation determined from base and factor dispersion through Eq. (20). The registered multiplicative arm instead uses the fixed unit exponent $f(a)^1$ and has no corresponding dispersion-matching step. Their comparison therefore changes both functional form and correction intensity. Second, α_s also settles the dimensional question: it carries demand units

per unit of the dimensionless factor, so the additive term is commensurate with d_a^{base} . Third, the normalization domain is $\mathcal{A}_s$, identical to that of the multiplicative form. Fourth, the final re-normalization conserves D_s and therefore D_r . Fifth, non-negativity holds due to the clipping $[\cdot]_+$ applied before re-normalization. One structural consequence follows immediately: a uniform base allocation has zero dispersion, hence $\alpha_s = 0$ and the additive correction degenerates to the base itself; Uniform-base additive entries therefore coincide with the uncorrected Uniform allocation.

3.8. Prior loss

The alternative to post-correction is to inject auxiliary information during GNN training rather than after inference. We term this the *prior-loss* mechanism because auxiliary information is incorporated during training, before inference, in contrast to post-correction, which operates after inference; the term does not imply a Bayesian prior over model parameters. A prior loss term converts the NTL or Proximity distribution into a target weight distribution and penalizes deviation via forward Kullback–Leibler (KL) divergence.

For each source s , the target distribution over RCI grid points is:

$$q_{s,a} = \frac{\log(1 + v(a))}{\sum_{a' \in \mathcal{A}_s \cap \text{RCI}} \log(1 + v(a'))} \quad (21)$$

where $v(a)$ is $\text{ntl}(a)$ or $\text{prox}(a)$. The prior loss is:

$$\mathcal{L}_{\text{prior}} = \frac{1}{|\mathcal{V}_s|} \sum_s \sum_{a \in \mathcal{A}_s} q_{s,a} \log\left(\frac{q_{s,a}}{w_{sa}}\right) \quad (22)$$

Three support-set consequences follow from the definitions of Section 3.1. First, $q_{s,a} = 0$ for every grid point outside RCI, and no smoothing redistributes target mass off this support: the corresponding terms of Eq. (22) vanish under the convention $0 \log 0 = 0$; the implementation realizes the exact zeros up to a 10^{-8} floor, whose measured gradient share is small in the typical sub-region but heavy-tailed, and whose pull is directed toward the floor itself (Appendix A.5). Second, the softmax of Eq. (6) runs over all of $\mathcal{A}_s$, so grid points outside RCI can still receive demand: the forward KL divergence penalizes under-allocation on the support of q and leaves the remaining weights unconstrained. Third, the prior and reconstruction objectives therefore compete only on $\mathcal{A}_s \cap \text{RCI}$, where their balance is set by the fixed weights λ of Eq. (23); off the support, the land-use reconstruction loss alone governs the allocation.

Forward KL divergence ($D_{\text{KL}}(q||w)$) is chosen over its reverse ($D_{\text{KL}}(w||q)$) for a practical reason: it penalizes the GNN for under-allocating demand to grid points that the auxiliary source marks as active, but does not penalize surplus allocation to grid points where the auxiliary signal is weak. In load disaggregation terms, the GNN must respect high-activity locations flagged by NTL or Proximity, while remaining free to assign demand elsewhere based on land-use or other learned features. Reverse KL divergence would have the opposite effect, concentrating demand on a few dominant grid points and suppressing allocation elsewhere. This is undesirable when the auxiliary source is noisy or incomplete.

The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{landuse}} + \lambda_{\text{ntl}} \mathcal{L}_{\text{ntl-prior}} + \lambda_{\text{prox}} \mathcal{L}_{\text{prox-prior}} \quad (23)$$

Unlike post-correction, the prior-loss mechanism lets the GNN decide how much of the auxiliary signal to absorb, without introducing a second normalization at inference time. The trade-off is retraining: each prior configuration requires a full training run, whereas post-correction can be applied to any pre-trained model without additional computation.

3.9. Summary of evaluated configurations

Table 1 lists the 22 configurations evaluated in this paper, classified along the three design axes, with demand weighting and auxiliary-information integration are systematically varied. The spatial partitioning axis is fixed to Voronoi, the current industry standard, and while alternative partitioning methods exist, fixing this axis isolates the effects of the other two. Uniform-base additive configurations coincide with Uni by construction ($\alpha_s = 0$, Section 3.7) and are therefore not listed as separate columns; the feature-fusion control (GNNfusion) enters the matrix through its own integration path, and its definition accompanies the main results (Section 5.1).

4. Experimental design

Two main study areas are evaluated in parallel: 16 ITL-2 zones in Great Britain (1,891 primary substations) and the Ausgrid distribution area in New South Wales, Australia (143 zone substations, financial year 2008–09). Each region is represented by approximately 50,000 allocation grid points assigned to substations through a Voronoi partition. The Börde district in Germany contributes one descriptive configuration set (13 substations, 34 municipalities) and a common-topology downstream sensitivity example; it is not an independent case-study validation.

4.1. Data sources and processing

Table 2 summarizes the datasets used in the three study areas. The German case follows the same processing pipeline as the British one with two differences: the projected coordinate reference system (EPSG:25832 versus EPSG:27700) and the nighttime-light compositing window (January 2022–December 2023 versus January 2021–December 2022), the latter matching the more recent vintage of the German demand data. For Britain, population enters as mid-2023 estimates and Gross Value Added (GVA) as 2022 values against the 2021–22 demand reporting year; the Australian case aligns all inputs to FY2009.

Land-use classification aggregates OpenStreetMap tags into five functional classes: residential, commercial, industrial, agricultural, and other. For each grid point a , a land-use proportion vector $\mathbf{p}_a = (p_a^{\text{res}}, p_a^{\text{com}}, p_a^{\text{ind}}, p_a^{\text{agr}}, p_a^{\text{oth}}) \in \mathbb{R}^5$ is obtained by a spatial join. This vector feeds both the GPM weighting rules and the GNN node feature matrix.

Nighttime-light radiance is taken from the NOAA VIIRS Day/Night Band (DNB) monthly composites. The median of all cloud-free observations between January 2021 and December 2022 is used, matching the 2021–2022 reporting year of the British substation demand data [38]. The `avg_rad` band records mean radiance in nanowatts $\text{cm}^{-2} \text{sr}^{-1}$; the observed dynamic range across the British study area spans 0–267. The native pixel size is approximately 500 m. Each grid point receives an NTL value by bilinear interpolation; grid-point values within a Voronoi cell are then aggregated into the NTL correction factor defined in Section 3.

4.2. Cross-validation and evaluation protocol

The British main case uses a four-fold spatial cross-validation scheme. The 16 ITL-2 regions are split into four groups of four; in each fold, three groups (12 regions) form the training set, and the remaining group (4 regions) is held out for testing. This procedure runs under three random seeds (42, 123, 456) controlling both parameter initialization and fold assignment, producing $4 \times 3 = 12$ evaluation runs per GNN-based configuration.

Although spatial autocorrelation exists between adjacent ITL-2 regions, each region is modeled as a separate graph: sub-region and grid-point nodes belong to exactly one region, and no edge or message passing crosses region boundaries. Encoder parameters are shared across all training-set graphs within a fold, but the test-region graph is withheld during training, so no substation-level information from the test region can leak into the learned weights.

Table 1

Configuration matrix. All configurations use Voronoi spatial partitioning. N = NTL, P = Proximity ($\gamma = 2$), NP = both; Post. = multiplicative post-correction, Add. = additive post-correction, Prior. = in-training prior loss, Feat. = feature-level fusion (auxiliary values appended to the grid-point input features). Within each group, configurations are ordered by increasing auxiliary information.

	Uniform base			GPM base						GNN base										GNN fusion																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																				
	Uniform			UniNP			GPM			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			GPMaddNP			GPMaddN			GPMaddP			

Table 2

Data sources for the three study areas. Great Britain and Germany share one processing pipeline; the Australian sources are aligned to the FY2009 study period (Section 4.2).

Data	Resolution	Role
<i>Great Britain and Germany</i>		
Administrative boundaries [35–37]	—	Region delineation
Substation locations & demand [38–40]	Point	Substation evaluation target
Land use [41]	~100 tags → 5 classes	GPM weights; GNN features
NTL [42]	~500 m	NTL correction factor
Economic output [43,44]	ITL-3 / NUTS-3	GNN reconstruction loss
Population [45]	ITL-3	GNN reconstruction loss
<i>Australia (Ausgrid, FY2009)</i>		
Statistical boundaries [46]	—	Region delineation
Substation load & locations [41,47,48]	Point, 15-min	Ground truth
Land use [49]	50 m → 5 classes	GPM weights; GNN features
NTL [50]	~1 km	NTL correction factor
Economic attributes [51–53]	SA2/SA3	GNN reconstruction loss

All configurations are reported as the 16-region mean $\pm$ inter-region standard deviation. For GNN-based configurations, per-region metric values are first averaged over the three seeds before regional aggregation; the reported $\pm$ therefore reflects spatial heterogeneity, not training randomness. Static configurations (Uniform, GPM, and their post-corrected configurations) are deterministic and require no seed averaging.

The second main case is the Ausgrid distribution network in New South Wales, Australia, covering the Sydney, Central Coast, and Hunter service areas. Ground truth comprises 143 zone substations with usable metered records for the 2008–09 financial year (FY2009); the target quantity is each substation's annual peak demand, taken as the maximum valid 15-minute reading of the year without extrapolation, and an annual-energy variant of the entire evaluation is used as a sensitivity check. Demand is metered in MW (in Britain, MVA); each case is internally consistent, and cross-case comparisons are restricted to the direction of mechanism effects.

The spatial hierarchy mirrors the British setup. Twelve SA4 statistical areas are the study regions (independent graphs and inference units, the counterparts of the ITL-2 zones), and the 34 substation-bearing SA3 areas are the sub-regions with known aggregate demand, each represented by one sub-region node; a thirteenth SA4 is excluded because its only FY2009 record lacks a usable measurement. Grid points are generated by the same rule as in Britain (target 50,000 per region), and graph construction, encoder architecture, and all training hyperparameters are carried over unchanged. The trained configurations mirror the British set in full: the self-supervised base, all three prior-loss configurations, and the feature-fusion configuration are each trained from scratch under this protocol (the single-prior and fusion runs postdate the comparison registry and report raw permutation p -values, Section 4.3, except the GNNpriorN configuration, which also enters the nine-comparison family and whose Holm and raw values coincide); the full post-correction matrix is evaluated exactly as in Britain: both mechanisms, all three signals, on all three bases. Cross-validation uses four folds of three regions each over the 12 regions, under the same three seeds.

Two inputs change source with the study period, and the switches are disclosed explicitly. Land-use proportions are derived from the Catchment Scale Land Use of Australia raster (2015 release, New South Wales mapping largely of 2007 vintage), whose ALUM class codes are reclassified into the same five classes used in Britain, in place of OpenStreetMap tags; nighttime light uses the DMSP-OLS stable-lights band (F16 satellite, median of the 2008 and 2009 annual composites, ~1 km resolution), as VIIRS observations are not available for FY2009. Both switches keep the inputs contemporaneous with the FY2009

demand data, and replication of the mechanism findings under different land-use and radiance sources indicates that these findings do not hinge on a particular data product. Sub-region attributes use the 2009 estimated resident population and mean personal income for 2008–09, with sectoral shares proxied by the 2011 Census industry-of-employment structure, a stated substitute for the British GVA shares.

With 12 inference units and an average of about 12 substations per region (Britain: 16 units, ~118 substations per region), region-level metrics are noisier and the power for detecting heterogeneous effects is lower; the adapted inference protocol is given in Section 4.3. Five substations flagged by a day–night consumption-ratio screen as potentially affected by rooftop photovoltaics are retained in the main analysis and removed in a sensitivity check; no substation shows negative net load in FY2009.

The German descriptive case includes all 13 substations and 34 municipalities in a single train-and-evaluate pass using the same three seeds. Results are reported as the three-seed mean $\pm$ standard deviation. Given the small sample size, no statistical inference is attempted.

The German target is the provider field “Last aktuell in MW” from the retrieved `v_epa_uw_last` snapshot [40]. It is a current-load record rather than an interval-metered annual peak series. All 13 retained values occur in 2.5 MW increments, so the source does not document variation below that reporting step.

The German case study is not treated as a third statistical sample. It supplies one descriptive set of configuration-level direction checks and one common-topology sensitivity analysis. The case cannot establish cross-case transfer, and its synthetic network does not validate absolute planning outcomes.

To assess the influence of individual measured substations on the descriptive metrics, a jackknife leave-one-out (LOO) analysis removes each of the 13 substations from the metric calculation and recomputes the correlation on the remaining 12. The trained predictions remain fixed; this is an influence/stability diagnostic, not a retraining-based test of overfitting or train–test circularity.

To test how alternative demand inputs change a downstream calculation under common assumptions, a 14-bus synthetic 110 kV network is constructed from the substation coordinates of the 13 Börde primary substations. A minimum spanning tree, computed over the pairwise distance matrix of the 13 substations and the centroid slack bus introduced below (14 nodes), defines the 13 edges that connect all nodes into a radial topology. Each edge is parameterized with the standard German 110 kV overhead-conductor type 149-AL1/24-ST1 A. A fourteenth bus, representing the connection to the external high-voltage transmission network, is placed at the geographic centroid of the 13 substation coordinates and designated as the slack bus at $v = 1.02$ p.u. AC power flow is solved with *pandapower* [54]. The minimum-spanning-tree topology and the centroid-positioned slack bus are spatial approximations that do not reproduce the actual Avacon Netz 110 kV network; all quantities derived from this model quantify relative differences among disaggregation configurations under identical network assumptions, and absolute loading values should not be read as statements about the real Börde network. Fig. 2 shows the resulting topology overlaid on the 34 municipality boundaries.

Four primary metrics are computed at the substation level within each region r ; WAPE and mean-normalized RMSE are reported as supplementary scale-free errors. The root-mean-square error,

$$\text{RMSE}_r = \sqrt{\frac{1}{|T_r|} \sum_{j \in T_r} (\hat{d}_j - d_j^{\text{actual}})^2}, \quad (24)$$

measures the absolute accuracy of disaggregated demand. The mean absolute error,

$$\text{MAE}_r = \frac{1}{|T_r|} \sum_{j \in T_r} |\hat{d}_j - d_j^{\text{actual}}|, \quad (25)$$

Simplified 110 kV Network (MST approximation)

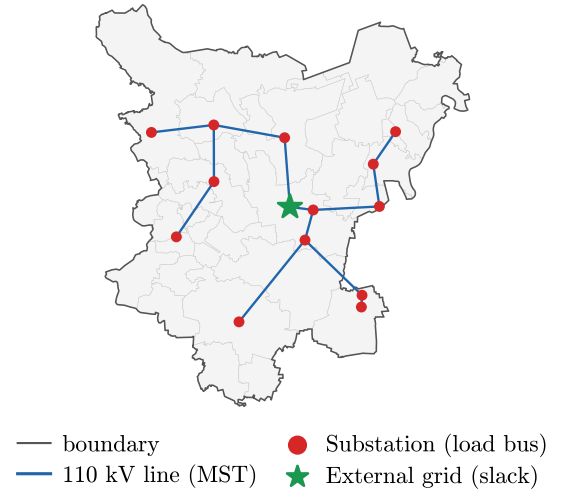


Fig. 2. Simplified 110 kV network for the Börde district used for downstream power-flow validation. Red circles represent the 13 primary substations; the green star marks the slack bus, positioned at the centroid, representing the external network connection. Blue edges are the 13 minimum spanning tree connections joining the 14 nodes (the 13 substations and the slack bus). Gray polygons show the 34 municipalities (Gemeinde) boundaries that define the demand aggregation units.

gives the average absolute deviation and is less sensitive to outliers than Root Mean Square Error (RMSE). The Pearson correlation coefficient,

$$\text{Corr}_r = \frac{\sum_j (\hat{d}_j - \bar{\hat{d}}) (d_j^{\text{actual}} - \bar{d}^{\text{actual}})}{\sqrt{\sum_j (\hat{d}_j - \bar{\hat{d}})^2} \sqrt{\sum_j (d_j^{\text{actual}} - \bar{d}^{\text{actual}})^2}}, \quad (26)$$

captures the linear association between predicted and observed substation demands.

Predictive R^2 compares squared prediction error with the within-region mean-demand reference:

$$R_{\text{pred},r}^2 = 1 - \frac{\sum_{j \in T_r} (\hat{d}_j - d_j^{\text{actual}})^2}{\sum_{j \in T_r} (d_j^{\text{actual}} - \bar{d}_r^{\text{actual}})^2} = 1 - \frac{\text{SSE}_r}{\text{SST}_r}. \quad (27)$$

Unlike squared correlation, this value can be negative. Pooled predictive R^2 uses $1 - \sum_r \text{SSE}_r / \sum_r \text{SST}_r$. WAPE and mean-normalized RMSE are $\sum_j |d_j^{\text{actual}} - \hat{d}_j| / \sum_j |d_j^{\text{actual}}|$ and $\text{RMSE}_r / |d^{\text{actual}}|_r$, respectively.

Three references separate different questions. EqualStation predicts D_r/n_r and is the exact zero point of within-region predictive R^2 . Direct EqualSubregion predicts D_s/n_s at the substation output and does not traverse the grid/Voronoi pipeline. EqualSubregionGrid distributes each D_s uniformly across the active substation-cell intersections, shares each cell's mass over its grid points, conserves every D_s , and then uses the frozen Voronoi aggregation, which is held fixed across all reported comparisons. It is therefore a pipeline-matched grid-level reference, but not a unique fair reference. The construction is reverse-defined from the substation cells, and a different mass-conserving grid construction can reproduce Direct EqualSubregion by zeroing intersections whose substation lies outside the source sub-region. That alternative zeroes 27.26% of British and 27.88% of Australian sub-regional mass and is treated as a modeling observation rather than a preferred demand surface.

4.3. Statistical inference protocol

All statistical claims in this paper follow one protocol, stated here once and only referenced in the results sections. The unit of inference

Table 3

Controlled two-axis evaluation on the two main study areas. Each configuration is a design cell rather than a ranked candidate. EqualSubregionGrid is the pipeline-matched grid-level reference: it uses the same grid support, known D_{ij} , arithmetic conservation, frozen Voronoi assignment, and grid-to-substation aggregation as the evaluated configurations. It is not a unique fair reference; the two pipeline-external references are documented in [Appendix D.5](#). Britain reports the 16-region mean $\pm$ inter-region standard deviation and Australia the corresponding 12-region values; GNN values are seed-averaged first. Predictive- R^2 cells additionally report the pooled value in parentheses, using $1 - \sum \text{SSE}_r / \sum \text{SST}_r$ with every SST_r centered on its own region mean. By variance decomposition, replacing those local means with one global mean can only increase the denominator, so this is the conservative pooled construction. Its denominator is concentrated: London supplies 47.3% of British SST and Sydney City and Inner South supply 52.2% of Australian SST. Demand is measured in MVA (GB) and MW (AU).

Configuration	RMSE		MAE		Corr		Predictive R^2	
	GB	AU	GB	AU	GB	AU	GB	AU
<i>Pipeline-matched grid reference</i>								
Grid Base	8.59 $\pm$ 2.98	21.74 $\pm$ 8.82	6.50 $\pm$ 2.00	17.78 $\pm$ 7.10	0.143 $\pm$ 0.141	0.048 $\pm$ 0.472	-0.493 $\pm$ 0.379 (-0.142)	-1.137 $\pm$ 1.789 (-0.368)
<i>Static bases</i>								
Uni	13.10 $\pm$ 3.70	28.29 $\pm$ 13.55	9.23 $\pm$ 2.21	20.93 $\pm$ 9.93	0.009 $\pm$ 0.178	0.038 $\pm$ 0.422	-3.626 $\pm$ 3.984 (-1.351)	-3.582 $\pm$ 5.494 (-1.735)
UniN	11.16 $\pm$ 3.29	26.47 $\pm$ 11.96	7.96 $\pm$ 2.02	19.77 $\pm$ 9.30	0.094 $\pm$ 0.190	0.055 $\pm$ 0.427	-2.245 $\pm$ 2.638 (-0.776)	-2.599 $\pm$ 3.343 (-1.331)
UniP	8.98 $\pm$ 2.96	18.53 $\pm$ 10.18	6.71 $\pm$ 1.90	14.76 $\pm$ 8.08	0.119 $\pm$ 0.194	0.166 $\pm$ 0.459	-0.794 $\pm$ 0.981 (-0.253)	-0.433 $\pm$ 0.909 (-0.261)
UniNP	7.45 $\pm$ 2.79	17.62 $\pm$ 10.41	5.50 $\pm$ 1.68	14.10 $\pm$ 8.23	0.319 $\pm$ 0.194	0.224 $\pm$ 0.445	-0.183 $\pm$ 0.547 (+0.074)	-0.245 $\pm$ 0.803 (-0.176)
GP	12.39 $\pm$ 3.36	29.11 $\pm$ 13.36	8.74 $\pm$ 2.11	21.45 $\pm$ 10.13	0.035 $\pm$ 0.199	0.030 $\pm$ 0.431	-3.243 $\pm$ 3.556 (-0.986)	-3.586 $\pm$ 4.872 (-1.752)
GPMpostN	10.54 $\pm$ 2.80	27.26 $\pm$ 12.10	7.56 $\pm$ 1.84	20.36 $\pm$ 9.70	0.128 $\pm$ 0.208	0.047 $\pm$ 0.436	-1.922 $\pm$ 2.223 (-0.477)	-2.632 $\pm$ 2.893 (-1.360)
GPMpostP	8.54 $\pm$ 2.50	19.29 $\pm$ 11.04	6.45 $\pm$ 1.72	15.23 $\pm$ 9.00	0.152 $\pm$ 0.179	0.143 $\pm$ 0.476	-0.597 $\pm$ 0.698 (-0.066)	-0.508 $\pm$ 0.890 (-0.310)
GPMpostNP	7.31 $\pm$ 2.40	18.44 $\pm$ 11.36	5.44 $\pm$ 1.50	14.61 $\pm$ 9.22	0.332 $\pm$ 0.173	0.205 $\pm$ 0.465	-0.129 $\pm$ 0.400 (+0.174)	-0.343 $\pm$ 0.865 (-0.236)
GPmaddN	10.39 $\pm$ 3.02	27.02 $\pm$ 13.54	7.52 $\pm$ 2.03	20.36 $\pm$ 10.45	0.123 $\pm$ 0.208	0.075 $\pm$ 0.468	-1.653 $\pm$ 1.844 (-0.419)	-2.879 $\pm$ 3.887 (-1.446)
GPmaddP	10.65 $\pm$ 3.10	26.70 $\pm$ 12.32	7.70 $\pm$ 2.13	19.79 $\pm$ 9.25	0.069 $\pm$ 0.209	0.062 $\pm$ 0.450	-1.789 $\pm$ 1.942 (-0.478)	-2.988 $\pm$ 4.373 (-1.321)
GPmaddNP	10.60 $\pm$ 3.17	26.65 $\pm$ 12.33	7.60 $\pm$ 2.15	19.75 $\pm$ 9.25	0.086 $\pm$ 0.211	0.064 $\pm$ 0.450	-1.791 $\pm$ 2.026 (-0.459)	-2.972 $\pm$ 4.362 (-1.315)
<i>GNN base</i>								
GNN	9.27 $\pm$ 2.61	40.97 $\pm$ 40.91	6.67 $\pm$ 1.76	28.04 $\pm$ 18.07	0.298 $\pm$ 0.201	0.009 $\pm$ 0.431	-1.068 $\pm$ 1.211 (-0.235)	-5.017 $\pm$ 4.972 (-10.161)
GNNpostN	9.45 $\pm$ 2.55	39.65 $\pm$ 41.08	6.58 $\pm$ 1.64	27.18 $\pm$ 18.13	0.365 $\pm$ 0.187	0.024 $\pm$ 0.437	-1.048 $\pm$ 0.931 (-0.276)	-4.393 $\pm$ 4.388 (-9.901)
GNNpostP	9.42 $\pm$ 2.33	29.23 $\pm$ 37.52	6.59 $\pm$ 1.52	19.50 $\pm$ 17.46	0.349 $\pm$ 0.166	0.163 $\pm$ 0.442	-0.987 $\pm$ 0.794 (-0.223)	-1.834 $\pm$ 3.444 (-6.889)
GNNpostNP	11.20 $\pm$ 2.64	28.74 $\pm$ 37.65	7.64 $\pm$ 1.66	19.09 $\pm$ 17.56	0.348 $\pm$ 0.160	0.190 $\pm$ 0.423	-1.841 $\pm$ 1.157 (-0.667)	-1.705 $\pm$ 3.447 (-6.837)
GNNaddN	8.82 $\pm$ 2.63	34.86 $\pm$ 43.59	6.35 $\pm$ 1.84	24.57 $\pm$ 20.75	0.452 $\pm$ 0.184	0.218 $\pm$ 0.390	-0.727 $\pm$ 0.735 (-0.156)	-4.268 $\pm$ 7.875 (-9.311)
GNNaddP	7.03 $\pm$ 2.27	19.97 $\pm$ 13.26	5.20 $\pm$ 1.46	15.20 $\pm$ 8.47	0.405 $\pm$ 0.164	0.166 $\pm$ 0.490	-0.034 $\pm$ 0.331 (+0.246)	-0.691 $\pm$ 1.400 (-0.707)
GNNaddNP	7.34 $\pm$ 2.14	19.58 $\pm$ 13.32	5.37 $\pm$ 1.41	14.94 $\pm$ 8.48	0.448 $\pm$ 0.188	0.194 $\pm$ 0.486	-0.189 $\pm$ 0.475 (+0.193)	-0.608 $\pm$ 1.306 (-0.668)
GNNpriorN	9.19 $\pm$ 3.02	41.98 $\pm$ 47.57	6.65 $\pm$ 1.96	28.19 $\pm$ 20.87	0.335 $\pm$ 0.173	0.031 $\pm$ 0.442	-0.885 $\pm$ 0.886 (-0.279)	-5.218 $\pm$ 6.116 (-12.723)
GNNpriorP	9.14 $\pm$ 2.89	40.93 $\pm$ 43.87	6.60 $\pm$ 1.88	27.69 $\pm$ 19.57	0.335 $\pm$ 0.176	0.019 $\pm$ 0.445	-0.885 $\pm$ 0.895 (-0.240)	-4.947 $\pm$ 5.487 (-11.284)
GNNpriorNP	9.03 $\pm$ 2.88	40.97 $\pm$ 42.94	6.51 $\pm$ 1.86	27.93 $\pm$ 18.91	0.342 $\pm$ 0.179	0.015 $\pm$ 0.438	-0.846 $\pm$ 0.881 (-0.224)	-4.849 $\pm$ 5.055 (-10.859)
GNNfusion	11.26 $\pm$ 3.26	33.30 $\pm$ 30.77	8.20 $\pm$ 2.41	24.00 $\pm$ 15.46	0.363 $\pm$ 0.178	0.103 $\pm$ 0.445	-1.894 $\pm$ 1.262 (-0.962)	-3.284 $\pm$ 3.131 (-6.653)

is the study region. For each planned comparison of two configurations, per-region metric values of GNN-based configurations are first averaged over the three seeds; the paired per-region differences then enter an exact two-sided sign-flip permutation test, enumerating all $2^{16} = 65,536$ sign assignments for the 16 British regions and all $2^{12} = 4096$ for the 12 Australian regions. Comparisons between two deterministic static configurations use the same test without seed averaging. This protocol replaces a Wilcoxon signed-rank procedure whose reported p -value was the median across seeds and left the unit of inference ambiguous.

Seed averaging is a design choice, not a hidden aggregation: per-seed permutation p -values are reported in full in [Appendix D](#), and where a single value per seed triple is quoted alongside them, the Cauchy combination test [55] is used, which remains valid under the dependence between seeds; median-across-seeds rule is abandoned. Confidence intervals for paired differences come from a region-level paired bootstrap ($B = 10^4$, percentile); resampling is over regions only, seeds are never resampled, and the bootstrap is not used to derive p -values. As a robustness check, a linear mixed model with crossed random intercepts for region and seed re-estimates each contrast.

All planned comparisons were enumerated in a registry before testing, and Holm's correction is applied within each metric family; significance statements refer to two-sided Holm-corrected $p < 0.05$. Because study regions are spatially dependent, Moran's I is computed for every British registered comparison on the paired per-region differences, using queen contiguity as the main spatial weight matrix (no region is isolated) with k -nearest-neighbors ($k = 3$) weights as a supplement. Where this diagnostic flags significant spatial autocorrelation, the primary p -value of that comparison is replaced by a spatial block permutation that flips signs at the level of contiguous region blocks; diagnostic values

and block definitions appear in [Appendix D](#). The same queen-contiguity and $k = 3$ diagnostics are applied to the 27 Australian registered comparison-metric pairs. Three correlation contrasts are flagged by the queen diagnostic; their contiguous-block permutation values are 0.094, 0.125, and 0.148, so no Australian RMSE headline or significance verdict changes. For the Australian sensitivity analysis at the SA3 level (34 units), exact enumeration is impractical and a Monte-Carlo sign-flip test with 10^5 draws with a fixed seed is used instead.

5. Results

The results proceed from aggregate comparison to integration-mechanism diagnostics and scope assessment. [Section 5.1](#) compares aggregate performance and statistical significance across the full configuration matrix across both main study areas. The RMSE-Pearson Correlation Coefficient (Corr) decoupling phenomenon is documented in [Section 5.2](#); [Section 5.3](#) quantifies interaction effects across integration routes; and [Section 5.4](#) tests candidate explanations with diagnostic controls without claiming factorial causal isolation. [Section 5.5](#) then tests the conditional-benefit criterion of [Section 3.5](#) across all three study cases, and [Sections 5.6](#) and [5.7](#) check the findings against intensity sweeps and concentration rescaling. The closing subsections assess regional heterogeneity ([Section 5.8](#)), an internal direction diagnostic against matching always-apply rates ([Section 5.9](#)), and the descriptive German case ([Section 5.10](#)).

5.1. Aggregate performance and statistical significance

[Table 3](#) reports RMSE, Mean Absolute Error (MAE), Corr, and predictive R^2 for the 22 evaluated configurations and the pipeline-matched

EqualSubregionGrid Base. Each metric is first computed within a region; GNN results are seed-averaged before regional aggregation. Predictive- R^2 cells also report the within-region pooled value in parentheses.

EqualSubregionGrid Base makes the weighting-layer result explicit. In Britain its pooled predictive R^2 is -0.142 , compared with -1.351 , -0.986 , and -0.235 for the uncorrected Uni, GPM, and GNN bases. Its region-mean value is -0.493 , compared with -1.068 for GNN. Uni and EqualSubregionGrid use the same known totals and both conserve demand; the difference is that Uni allocates by grid-cell area, whereas EqualSubregionGrid divides mass by the number of active substation-cell intersections. Applying the additive NP correction to the same Grid construction raises pooled predictive R^2 from -0.142 to -0.009 , so this within-base contrast, rather than the base-only comparison, shows the contribution of the correction step.

Within each main study case, 15 of the 19 corrected British configurations and 18 of the 19 Australian configurations reduce region-mean RMSE relative to their uncorrected base; 14 do so in both. A stricter comparison uses the highest pooled predictive R^2 among the seven Grid configurations, attained by Grid addNP (-0.009 in Britain and -0.157 in Australia). Within the preregistered comparison set, UniNP and GPMpostNP exceed the British Grid value, while no configuration exceeds the Australian value. The two post-registry GNN additive configurations also exceed Grid addNP in Britain, but they are disclosed separately.

The two estimands give different Australian readings. EqualSubregionGridaddNP reaches a pooled predictive $R^2 = -0.157$, above UniNP at -0.176 . Under region-equal averaging, however, UniNP is higher (-0.245 versus -0.954). Both results remain negative. In Britain, GNNaddP and GPMpostNP remain statistically indistinguishable under region-equal RMSE (Holm $p = 0.163$). Their pooled predictive- R^2 values are 0.246 and 0.174 ; the descriptive paired region-bootstrap interval for the 0.072 difference is $[0.027, 0.115]$. This post-registry interval does not replace the registered region-equal comparison. Against EqualStation, GNNaddP and GPMpostNP each have lower RMSE in 7/16 regions and UniNP in 10/16; against the direct EqualSubregion, the counts are 5/16, 3/16, and 3/16, respectively.

The pipeline-external references are documented in Appendix D.5. Direct EqualSubregion divides each D_s equally among its substations without traversing the grid/Voronoi route and attains a pooled predictive R^2 of 0.248 in Britain and 0.194 in Australia. The grid route therefore does not add any absolute accuracy over substation-level equal division; this study instead measures correction direction within a fixed route.

The pooled denominator is concentrated in both of the main cases. London supplies 47.3% of British SST; after it is removed, GNNaddP changes from 0.246 to 0.112 and GPMpostNP from 0.174 to 0.043 . Sydney City and Inner South supply 52.2% of Australian SST; removing it changes UniNP from -0.176 to $+0.058$. Region-mean and pooled values are therefore reported together.

The negative British region-mean values partly reflect denominator heterogeneity. Across the 16 regions, the correlation between within-region demand standard deviation and predictive R^2 is 0.734 for GNNaddP and 0.711 for GPMpostNP. The seven positive- R^2 GNNaddP regions have mean demand standard deviation 10.04 , compared with 5.45 elsewhere. This explanation does not apply to Australia: the corresponding correlation is 0.152 , the group means are 16.95 and 16.69 , and RMSE varies more across regions than demand standard deviation. The Australian negative values therefore remain an accuracy failure rather than a denominator artifact.

WAPE places the errors relative to the observed load and ranges from 38.5–68.5% in Britain and 34.0–66.1% in Australia. In Britain, GNNaddP (38.5%), GNNaddNP (39.8%), GPMpostNP (40.4%), and UniNP (40.7%) are below EqualStation (40.8%). Australian UniNP is likewise lower than EqualStation (34.0% versus 34.6%), while the direct EqualSubregion is lower still at 27.5%. WAPE can therefore favor configurations whose predictive R^2 remains negative. Full WAPE and mean-normalized RMSE

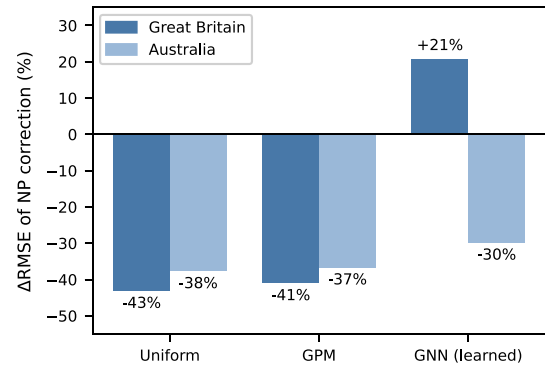


Fig. 3. The sign flip at a glance: relative RMSE change of the multiplicative NTL+Proximity post-correction for each base allocation (16-region mean for Great Britain, 12-region mean for Australia; Table 3). Both static bases improve by 37–43% in both study cases; the learned base is harmed in Britain (+21%) yet helped in Australia (−30%). This configuration-level reversal motivates the empirical test of the conditional-benefit criterion in Section 5.5; it is not treated as an independent study-case validation.

results are provided in Appendix D.5; MAPE is not used because five British regions contain six zero-demand substations.

The feature-fusion control GNNfusion appends the NTL and Proximity values to the grid-point feature vector (seven input features instead of five) and retrain from scratch under an otherwise identical protocol; it tests whether feature-level injection can substitute for either correction mechanism. The two study cases answer in opposite directions: the British RMSE rises from 9.27 to 11.26 , while the Australian one falls from 40.97 to 33.30 (−18.7%, exact sign-flip $p = 0.015$), showing the same base-state dependence that governs the explicit corrections (Section 5.5). Section 5.4 returns to this configuration with probing evidence on where the injected information ends up and what determines its effect.

Fig. 3 shows the configuration-dependent phenomenon examined in the remainder of the paper: the same multiplicative NP correction, applied to the three bases in the two main study cases.

Table 4 reports the nine planned comparisons on RMSE under the inference protocol of Section 4.3; none of the 27 British comparison-metric significance verdicts changes direction relative to a Wilcoxon-based alternative.

In Britain, six of the nine comparisons are statistically significant. They point to three conclusions.

In the British study case, GNN has an RMSE 3.12 lower than GPM (comparison #1, $p < 0.001$), showing that end-to-end learning on the same five-dimensional land-use features can replace hand-crafted weighting rules there. The Australian GNN base does not reproduce this effect (40.97 ± 40.91 , Corr 0.009), a failure consistent with the smaller training corpus (34 source sub-regions) and the older, re-sourced FY2009 inputs (Section 4.2).

The static correction chain, by contrast, exhibits layer-wise synergy. Proximity added to GPM yields $\Delta\text{RMSE} = -3.85$ (#6); NTL added on top of GPM + Proximity yields a further -1.24 (#7); Proximity added on top of GPM + NTL yields -3.23 (#5). All three comparisons remain significant under the protocol (Holm $p \leq 3.1 \times 10^{-4}$; where spatially flagged, block-permutation $p = 0.008$), so each correction layer contributes information that the others do not already capture.

For the GNN base, the pattern reverses. The combined post-correction is significantly antagonistic: GNNpostNP versus GNNpostP gives $\Delta\text{RMSE} = +1.78$ (Holm $p = 2.7 \times 10^{-4}$, block-permutation $p = 0.008$, #4). GPMpostNP has an RMSE 2.11 lower than the GNNpostP configuration (Holm $p = 2.7 \times 10^{-4}$, block-permutation $p = 0.008$, #9).

The three non-significant comparisons reinforce rather than weaken this reading. Single post-corrections on the GNN (#2, #3) show consistent RMSE deterioration ($+0.15$, $+0.19$) that does not reach

Table 4

Planned comparisons on RMSE: the same nine planned comparisons in both main study cases, tested by region-level exact two-sided sign-flip permutation on seed-averaged paired region differences (2^{16} enumeration over the 16 British regions, 2^{12} over the 12 Australian regions) and Holm-corrected within each nine-comparison metric family per study case (protocol in Section 4.3). Where Moran's I flags spatial autocorrelation, the block-permutation value replaces the primary p (British column block p ; 8 contiguous blocks, smallest attainable ≈ 0.008). Australian enumeration has the smallest raw two-sided $p = 2/2^{12} \approx 0.00049$ and a nine-comparison Holm floor of approximately 0.004; rows #2, #5, and #6 reach that displayed floor. Australian queen diagnostics flag three correlation contrasts, all non-significant after contiguous-block permutation ($p \geq 0.094$); full diagnostics appear in Appendix D. Positive Δ RMSE indicates deterioration.

#	Comparison	Britain				Australia			
		Δ RMSE	Holm p	Block p	Sig.	Δ RMSE	Holm p	Sig.	
1	GNN vs GPM	-3.12	4.9×10^{-4}	—	✓	+11.86	0.361	—	✓
2	GNNpostP vs GNN	+0.15	1.0	—	—	-11.74	0.004	✓	—
3	GNNpostN vs GNN	+0.19	1.0	0.562	—	-1.32	0.024	✓	—
4	GNNpostNP vs GNNpostP	+1.78	2.7×10^{-4}	0.008	✓	-0.48	0.336	—	—
5	GPMpostNP vs GPMpostN	-3.23	3.1×10^{-4}	0.008	✓	-8.83	0.004	✓	—
6	GPMpostP vs GPM	-3.85	2.7×10^{-4}	0.008	✓	-9.82	0.004	✓	—
7	GPMpostNP vs GPMpostP	-1.24	2.7×10^{-4}	—	✓	-0.85	0.023	✓	—
8	GNNpriorN vs GNN	-0.08	1.0	—	—	+1.01	0.956	—	—
9	GNNpostP vs GPMpostNP	+2.11	2.7×10^{-4}	0.008	✓	+10.79	0.336	—	—

significance. The prior-loss NTL configuration (#8) points in the beneficial direction (-0.08) but does not reach significance (Holm $p = 1.0$). On the RMSE metric, then, auxiliary information grafted onto the trained British GNN provides no significant benefit regardless of mechanism; the prior-loss route is simply the only one that avoids outright degradation. The Australian columns of Table 4 show the counterpart on a much weaker base (Corr 0.009): five of the six multiplicative comparisons improve RMSE significantly there (#2–3, #5–7), the sixth (#4) points the same beneficial way without reaching significance (-0.48 , Holm $p = 0.336$), and the single-prior configuration remains non-significant ($+1.01$, Holm $p = 0.956$). Of the two cross-pipeline comparisons, #1 reverses its British direction (the Australian GNN base sits $+11.86$ above GPM) and #9 retains it at a larger magnitude ($+10.79$), but neither reaches significance (Holm $p = 0.361$ and 0.336); Section 5.5 explains why the same mechanisms come down on opposite sides in the two study cases.

5.2. RMSE–Corr decoupling

Applying the same statistical tests to the Corr metric exposes a decoupling between calibration (absolute accuracy, measured by RMSE) and discrimination (measured by Corr) that is invisible to either metric alone. Table 5 contrasts the RMSE and Corr significance levels for four representative comparisons. For GNNpostN relative to GNN, RMSE exhibits a non-significant increase ($+0.19$, Holm $p = 1.0$), whereas Corr improves significantly ($+0.067$, from 0.298 to 0.365 , Holm $p = 5.5 \times 10^{-4}$). GNNpostP follows the same pattern: RMSE increases by $+0.15$ while Corr rises by $+0.051$.

The combined post-correction (GNNpostNP) further sharpens the pattern. Its Corr of 0.348 is statistically indistinguishable from that of GNNpostP (0.349 ; Holm $p = 1.0$), yet the RMSE deteriorates sharply to 11.20 . The third row of Table 5 confirms this asymmetry: the incremental NTL factor is significant for RMSE (antagonistic) but non-significant for Corr (no correlation gain).

The decoupling does not arise in the static pipeline. GPMpostNP improves both metrics at once (Corr 0.332 with RMSE 7.31), as shown in the last row of Table 5, where both British metrics are significant and improving. On the GPM base (Corr = 0.035), correction layers improve both calibration and discrimination in tandem. On the GNN base (Corr = 0.298), the same corrections improve discrimination while worsening calibration, a reversal of the pattern.

5.3. Synergy–antagonism asymmetry

To quantify how identical auxiliary information behaves under different integration mechanisms, the marginal RMSE change $\Delta_I^B =$

Table 5

RMSE–Corr decoupling: the same comparison can be significant on one metric and non-significant on the other. Significance under the protocol of Section 4.3, Holm-corrected within each metric family. Australian columns: the same comparisons under the Australian nine-comparison protocol of Table 4 (12 regions, exact 2^{12} sign-flip permutation, Holm within each nine-comparison metric family). On the Australian GNN base, whose residuals are large, the multiplicative corrections improve calibration and discrimination together, following the static-base pattern rather than the British learned-base decoupling; in the static-chain row (last), the Australian RMSE gain is significant while the Corr gain is not.

Comparison	Britain				Australia			
	RMSE		Corr		RMSE		Corr	
	Δ	Sig.	Δ	Sig.	Δ	Sig.	Δ	Sig.
GNNpostN vs GNN	+0.19	—	+0.067	✓	-1.32	✓	+0.015	✓
GNNpostP vs GNN	+0.15	—	+0.051	—	-11.74	✓	+0.154	✓
GNNpostNP vs GNNpostP	+1.78	✓	-0.001	—	-0.48	—	+0.027	—
GPMpostNP vs GPMpostN	-3.23	✓	+0.204	✓	-8.83	✓	+0.158	—

$\text{RMSE}(B + I) - \text{RMSE}(B)$ is computed for each information component I for each base B ; negative values indicate improvement. A positive marginal effect is termed *antagonistic* when it is statistically significant (two-sided Holm-corrected $p < 0.05$ under the protocol in Section 4.3); a positive but non-significant marginal effect is classified as *directionally adverse*.

The asymmetry across mechanisms is pronounced. On the static bases, NTL and Proximity are approximately additive: the combined GPM effect (-5.08) accounts for 89% of the sum of independent marginal effects ($-1.85 + (-3.85) = -5.70$), and the corresponding ratio on Uniform is 93% (-5.65 versus -6.06). The slight sub-additivity reflects partial overlap between NTL and Proximity in urbanized areas, though all 16 regions improve in every case (the registered static comparisons are all significant, Table 4 #5–7).

On the GNN post-correction route, single-factor corrections are directionally adverse but statistically non-significant (NTL alone: $+0.19$, $p = 1.0$; Proximity alone: $+0.15$, $p = 1.0$). As documented in Section 5.2, these corrections simultaneously produce a significant Corr improvement ($+0.067$); they are instances of metric decoupling rather than antagonism. Combining NTL and Proximity crosses the significance threshold and is antagonistic in the sense defined above: the deterioration ($+1.93$ RMSE, $+20.9\%$) is super-additive, far exceeding the sum

Table 6

Independent marginal RMSE effect of each auxiliary component in both main study cases, measured from the unmodified base: $\Delta = \text{RMSE}(\text{base} + I) - \text{RMSE}(\text{base})$, with percentages taken from the base of the same study case. Negative values denote improvement; positive values denote deterioration. Each row reports the effect of adding that component alone; the NTL + Prox row adds both simultaneously. Significance is assessed under the inference protocol of Section 4.3 (nine-comparison Holm families of Table 4 for the planned comparisons; exact sign-flip permutation for the remaining marginals). In the Australian panel every static-base marginal is a significant improvement, with the combined correction removing 37.7% (Uniform) and 36.7% (GPM) of base RMSE. The multiplicative GNN N-only and P-only marginals belong to the registered family (Holm $p = 0.024$ and 0.004), whereas the NP marginal is post-registry and reports its raw exact sign-flip value. Prior-loss marginals remain small and non-significant; the Australian NTL + Prox prior entry is -0.003 , shown as -0.00 at table precision.

Component	Uniform base	GPM base	GNN post-corr.	GNN prior loss
<i>Great Britain</i>				
NTL	-1.94 (-14.8%)	-1.85 (-15.0%)	+0.19 (+2.0%)	-0.08 (-0.9%)
Proximity	-4.12 (-31.5%)	-3.85 (-31.0%)	+0.15 (+1.6%)	-0.12 (-1.3%)
NTL + Prox	-5.65 (-43.1%)	-5.08 (-41.0%)	+1.93 (+20.9%)	-0.24 (-2.6%)
<i>Australia</i>				
NTL	-1.82 (-6.4%)	-1.85 (-6.4%)	-1.32 (-3.2%)	+1.01 (+2.5%)
Proximity	-9.75 (-34.5%)	-9.82 (-33.7%)	-11.74 (-28.7%)	-0.04 (-0.1%)
NTL + Prox	-10.66 (-37.7%)	-10.68 (-36.7%)	-12.23 (-29.8%)	-0.00 (-0.0%)

of the individual directionally adverse effects ($+0.19 + 0.15 = +0.34$), and the incremental contribution of adding NTL on top of Proximity correction is significant ($\Delta\text{RMSE} = +1.78$, Holm $p = 2.7 \times 10^{-4}$, block-permutation $p = 0.008$, Table 4 #4). Eq. (14) contains a positive cross-term for correlated factors only in its strict grid-unit/log-domain setting. Because this identity does not commute through Voronoi aggregation, the observed post-aggregation RMSE super-additivity is reported as an empirical interaction rather than as a consequence of that equation.

On the prior-loss route, the combined effect (-0.24 RMSE, -2.6%) points in the synergistic direction but does not reach significance (the registered single-prior comparison gives a Holm $p = 1.0$, Table 4 #8), and the improvement remains small.

The Australian panel of Table 6 puts the same decomposition on the other side: on the static bases the chain is strongly synergistic, the combined correction removing 37.7% of the Uniform base and 36.7% of the GPM base with every static marginal being significant; the three multiplicative GNN marginals are significant improvements as well (-1.32 to -12.23 ; raw sign-flip $p < 0.05$), while the prior-loss marginals stay between $+1.01$ and -0.003 and are non-significant. The two study cases share a wide multiplicative swing and a small prior-loss change. The sign of the multiplicative swing co-varies with the measured state of the base; Section 5.5 evaluates that association through the registered direction classifier.

The NTL and Proximity correction factors are moderately correlated across 122,441 RCI grid points (Pearson correlation 0.350, Spearman correlation 0.481, $p < 0.001$). Both factors encode aspects of urbanization intensity (NTL through nocturnal radiance, Proximity through substation siting density), and the strength of their overlap varies widely across regions: the per-region Spearman correlation spans from 0.241 (TLG2) to 0.685 (TLC2). Section 5.4 compares this association with diagnostic controls; it does not identify a unique causal component.

5.4. Integration-mechanism diagnostics

Three diagnostic experiments probe candidate explanations for post-correction antagonism on the GNN base. They constrain particular explanations but do not form a factorial causal design. Each follows the same $3 \times 4 = 12$ seed-fold evaluation protocol used in the main experiments. Table 7 collects the full numerical results for all three experiments and their single-factor variants, for both main study cases.

Removing the denominator of Eq. (11) and applying raw multiplicative scaling raises RMSE from 11.20 to 17.53 ($+56.5\%$; Table 7), because demand conservation is violated. On its own, this experiment cannot separate normalization from conservation: removing the denominator changes both at once. Proposition 1 supplies the separation analytically: the re-normalization multiplies each sub-region by a single positive constant, so the two variants differ exactly by conservation offsets, and the entire degradation of the no-re-normalization variant is a scale effect (confirmed to machine precision, Section 3.5). The second normalization is therefore a structural necessity that contributes no shape distortion of its own; double-normalization is not the cause of the antagonism.

If the antagonism stems from signal redundancy, i.e., the GNN having already encoded NTL and Proximity information so that the post-correction double-counts it, then replacing the real correction factors with spatially uncorrelated random noise should eliminate the effect. Random-noise multiplicative post-correction with re-normalization yields an RMSE = 9.59 ± 2.68 (Table 7), only marginally above the uncorrected GNN base (9.27). This is below the real NTL + Proximity multiplicative correction (11.20), showing that the chosen spatial factor produces a larger change than the random-factor control. The contrast does not identify which property of the real factor causes the difference.

The third experiment compares the specified multiplicative correction operator with the additive correction operator in Eqs. (19) and (20). The additive route changes centering, clipping, and re-normalization and also matches correction dispersion through the label-free, sub-region-specific scale α_s , whereas the multiplicative arm retains unit intensity. This is therefore a correction-procedure comparison that mixes functional-form and intensity differences rather than a factorial isolation of multiplication alone. Additive NTL + Proximity post-correction yields RMSE = 7.34 ± 2.14 and Corr = 0.448 ± 0.188 (Table 7), a large improvement over the multiplicative form (RMSE 11.20, Corr 0.348) and a value close to GPMpostNP (7.31).

Together, the controls rule out a simple double-normalization explanation and show that the tested additive and multiplicative correction procedures behave differently on the British learned base. Because the procedures differ in several transformations, they do not establish multiplication as the sole causal variable or imply that an additive correction cannot do harm in another setting. The result is reported as a correction-procedure-specific empirical contrast.

The Australian controls have the same directions as the British controls (Table 7, lower panel): random-noise correction changes the RMSE little (41.04 versus 40.97), and removing re-normalization worsens the dual-signal result ($28.74 \rightarrow 39.17$). These controls constrain two candidate explanations but do not isolate a universal causal component of the correction procedure.

At the representation layer, whether the GNN internally encodes NTL or Proximity information is tested with a probing suite that carries its own power checks rather than a single-dimension correlation probe. A nonlinear decoder (a multilayer perceptron on standardized 128-dimensional grid-point embeddings) is trained to predict the log-transformed, per-region standardized NTL and Proximity values and is evaluated out of region: probe training and testing never share a region (grouped four-fold splits), and random splits are reported only as an inflation contrast in Appendix D. Table 8 reports the suite. The positive control confirms probe power ($R^2 > 0.999$ when the inputs contain the targets), and the detectability check verifies signal recovery: on the feature-fusion configuration, whose inputs contain the signals, the probe recovers them ($R^2 \geq 0.97$; paired gain over base embeddings $+0.94$, sign-flip $p = 3.1 \times 10^{-5}$). Against these two calibrations, base embeddings yield $R^2 = 0.074$ (NTL) and 0.040 (Proximity), below even the raw five land-use features (0.357 and 0.136). The base GNN embeddings therefore show low out-of-region decodability of the auxiliary signals beyond what the land-use inputs imply; the full suite, CKA values, and the random-split inflation appear in Appendix D. The detectability check also answers the feature-fusion question raised in Section 5.1: the fusion configuration absorbs the auxiliary signals into its representation,

Table 7

Integration-mechanism diagnostics: experiments on the GNN base allocation, run in full on both main study cases (Britain: 16-region mean $\pm$ inter-region std; Australia: 12-region mean $\pm$ inter-region std; both under the $3 \times 4 = 12$ seed-fold protocol, seed-averaged before aggregation). “Mult.” = multiplicative post-correction with re-normalization; “Add.” = additive post-correction with re-normalization; “No-renorm” = multiplicative without re-normalization; “Random” = spatially uncorrelated random noise with re-normalization. Δ RMSE is relative to the uncorrected GNN base of the same study case. Together, the controls show that the tested correction procedures behave differently and rule out a simple double-normalization explanation; because they do not vary the functional form alone, they do not attribute the British result to multiplication as the sole causal variable. In both study cases, random correction changes RMSE little (+3.5% Britain, +0.2% Australia), while removing re-normalization worsens the dual-signal correction relative to its re-normalized counterpart (+56.5% and +36.3%).

Experiment	Correction form	RMSE	Δ RMSE	MAE	Corr
Great Britain					
<i>Controls</i>					
GNN base (no correction)	—	9.27 $\pm$ 2.61	—	6.67 $\pm$ 1.76	0.298 $\pm$ 0.201
GNN + NTL $\times$ Prox (Mult.)	Mult. + renorm	11.20 $\pm$ 2.64	+1.93 (+20.9%)	7.64 $\pm$ 1.66	0.348 $\pm$ 0.160
GNN + NTL only (Mult.)	Mult. + renorm	9.45 $\pm$ 2.55	+0.19 (+2.0%)	6.58 $\pm$ 1.64	0.365 $\pm$ 0.187
GNN + Prox only (Mult.)	Mult. + renorm	9.42 $\pm$ 2.33	+0.15 (+1.6%)	6.59 $\pm$ 1.52	0.349 $\pm$ 0.166
<i>Exp 1: Removing re-normalization</i>					
No-renorm NTL $\times$ Prox	Mult.	17.53 $\pm$ 6.75	+8.26 (+89.1%)	10.71 $\pm$ 3.34	0.357 $\pm$ 0.167
No-renorm NTL only	Mult.	10.37 $\pm$ 2.82	+1.10 (+11.9%)	7.10 $\pm$ 1.87	0.368 $\pm$ 0.187
No-renorm Prox only	Mult.	11.62 $\pm$ 4.20	+2.35 (+25.4%)	7.63 $\pm$ 2.22	0.357 $\pm$ 0.170
<i>Exp 2: Random-noise post-correction</i>					
Random noise (10 repeats avg.)	Mult. + renorm	9.59 $\pm$ 2.68	+0.33 (+3.5%)	6.88 $\pm$ 1.82	0.287 $\pm$ 0.199
<i>Exp 3: Additive post-correction</i>					
Additive NTL $\times$ Prox	Add. + renorm	7.34 $\pm$ 2.14	−1.92 (−20.8%)	5.37 $\pm$ 1.41	0.448 $\pm$ 0.188
Additive NTL only	Add. + renorm	8.82 $\pm$ 2.63	−0.45 (−4.8%)	6.35 $\pm$ 1.84	0.452 $\pm$ 0.184
Additive Prox only	Add. + renorm	7.03 $\pm$ 2.27	−2.24 (−24.1%)	5.20 $\pm$ 1.46	0.405 $\pm$ 0.164
Australia					
<i>Controls</i>					
GNN base (no correction)	—	40.97 $\pm$ 40.91	—	28.04 $\pm$ 18.07	0.009 $\pm$ 0.431
GNN + NTL $\times$ Prox (Mult.)	Mult. + renorm	28.74 $\pm$ 37.65	−12.23 (−29.8%)	19.09 $\pm$ 17.56	0.190 $\pm$ 0.423
GNN + NTL only (Mult.)	Mult. + renorm	39.65 $\pm$ 41.08	−1.32 (−3.2%)	27.18 $\pm$ 18.13	0.024 $\pm$ 0.437
GNN + Prox only (Mult.)	Mult. + renorm	29.23 $\pm$ 37.52	−11.74 (−28.7%)	19.50 $\pm$ 17.46	0.163 $\pm$ 0.442
<i>Exp 1: Removing re-normalization</i>					
No-renorm NTL $\times$ Prox	Mult.	39.17 $\pm$ 44.21	−1.80 (−4.4%)	24.66 $\pm$ 18.19	0.178 $\pm$ 0.408
No-renorm NTL only	Mult.	43.56 $\pm$ 48.03	+2.59 (+6.3%)	28.96 $\pm$ 19.57	0.029 $\pm$ 0.445
No-renorm Prox only	Mult.	33.74 $\pm$ 38.15	−7.23 (−17.6%)	22.19 $\pm$ 16.87	0.157 $\pm$ 0.427
<i>Exp 2: Random-noise post-correction</i>					
Random noise (10 repeats avg.)	Mult. + renorm	41.04 $\pm$ 40.86	+0.06 (+0.2%)	28.12 $\pm$ 18.07	0.010 $\pm$ 0.425
<i>Exp 3: Additive post-correction</i>					
Additive NTL $\times$ Prox	Add. + renorm	19.58 $\pm$ 13.32	−21.39 (−52.2%)	14.94 $\pm$ 8.48	0.194 $\pm$ 0.486
Additive NTL only	Add. + renorm	34.86 $\pm$ 43.59	−6.11 (−14.9%)	24.57 $\pm$ 20.75	0.218 $\pm$ 0.390
Additive Prox only	Add. + renorm	19.97 $\pm$ 13.26	−21.00 (−51.3%)	15.20 $\pm$ 8.47	0.166 $\pm$ 0.490

Table 8

Probing suite (key results): out-of-region R^2 of a nonlinear probe predicting the log-transformed, per-region standardized signals from each representation (region-blocked splits; pooled over 16 regions). Checks (i) and (ii) calibrate probe power and detectability; check (iii) is the claim under test; check (iv) is the information already present in the raw inputs. Full suite, CKA, and random-split inflation in Appendix D.

Representation	NTL	Prox
(i) Positive control (inputs contain targets)	0.99997	0.99998
(ii) Fusion embeddings (signals injected)	0.988	0.973
(iii) Baseline embeddings	0.074	0.040
(iv) Raw land-use features (5-dim)	0.357	0.136

yet in Britain its allocation is worse than that of the five-feature base (11.26 versus 9.27). Absorption alone therefore guarantees nothing; the training objective offers no gradient that rewards turning the signals

into better allocations (Section 3.3), and what the absorbed signals end up doing to the allocation depends on the state of the base, exactly as for the explicit corrections. The Australian retraining goes in the other direction on a weak base (pooled predictive $R^2 = -10.161$): the same seven-feature configuration lowers RMSE from 40.97 to 33.30 (−18.7%, raw exact sign-flip $p = 0.015$), while the British value rises from 9.27 to 11.26 (+21.5%). The Australian improvement does not cross EqualStation and is therefore interpreted as a relative change to a weak base, not adequate absolute accuracy. Like the Australian single-prior configurations, this run postdates the registry freeze, so its permutation p -values are raw rather than Holm-corrected. The feature-fusion outcomes add a representation-level example of configuration dependence. Section 5.5 formalizes only the multiplicative post-correction route; it does not supply a theorem for feature fusion. The probing suite itself is reported for Britain only: its checks rest on per-region frozen embeddings and the calibrated controls of the British diagnostic chain, and rebuilding that chain for the Australian configurations would be a new experimental campaign rather than an added column.

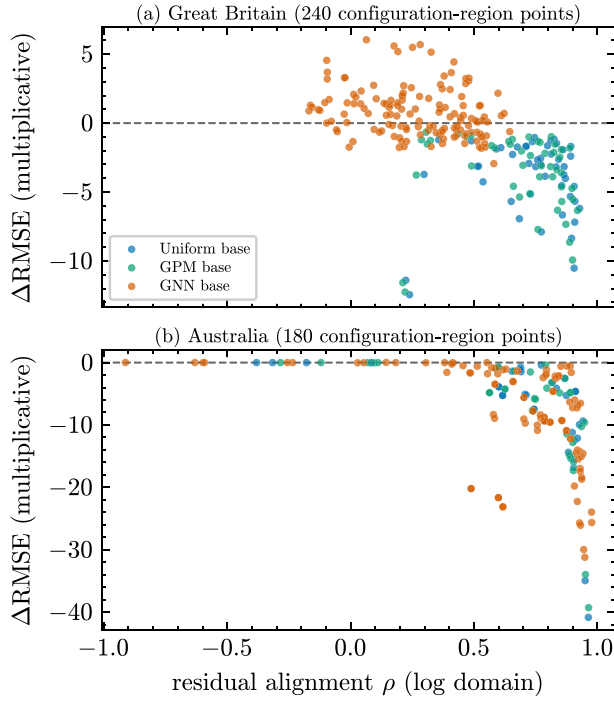


Fig. 4. Per-configuration ΔRMSE against substation-level factor-residual alignment for Britain (240 points nested in 16 regions, top) and Australia (180 points nested in 12 regions, bottom). The correction factor is the effective F_j derived after grid weighting, arithmetic sub-regional normalization, and Voronoi aggregation (Section 3.6). Negative ΔRMSE means that the correction helps. The points are repeated configurations within regions, not independent external validation cases.

5.5. Testing the conditional-benefit criterion

The exact result of Eq. (13) applies to geometric re-normalization and log-MSE. Here we first measure its ingredients at substations, then test how well the same fixed threshold classifies log-MSE and original-scale RMSE outcomes under the implemented arithmetic and Voronoi-aggregated pipeline. The empirical coordinates use the stated floor $\varepsilon = 10^{-6} \sum_j y_j$ and are therefore checked for floor sensitivity rather than being described as parameter-free. The pipeline result is an empirical prediction problem, not an extension of the if-and-only-if statement.

Factor-residual alignment is computed for every (region, base, signal) configuration after the grid-point correction is converted into the effective substation factor F_j of Eq. (17). We then compute the Pearson correlation between $\log F_j$ and the base log residual ratio, together with their population standard deviations. Centering $\log F_j$ does not change these coordinates (Section 3.6). On the static bases the factors track the residuals they are meant to fix: the median alignment is 0.767, computed over the 96 region-configuration pairs of the Uniform and GPM bases shown in the British panel of Fig. 4 (16 regions $\times$ 6 static configurations), and the multiplicative correction reduces RMSE in every one of them. On the GNN base the alignment collapses to a median of 0.299, computed over the 144 GNN-base pairs of the same panel, and only 38% of pairs improve. Alignment also predicts the size of the effect: across all 240 British pairs, the Spearman correlation between alignment and ΔRMSE is -0.691 ($p = 1.9 \times 10^{-35}$). The dispersion side of the criterion moves the same way: over the same 96 and 144 pairs, the median $\sigma_c/\sigma_{\text{res}}$ is 0.384 on the static bases against 0.518 on the GNN base (0.850 over the 48 dual-signal pairs), so the benefit threshold rises exactly where the alignment falls. Fig. 4 plots ΔRMSE against alignment for Britain and Australia: the static and learned bases separate cleanly, and the Australian points populate the high-alignment, negative- Δ side. Per-study-case confusion matrices and majority-class base rates are given

in Appendix D.7. The Australian learned base occupies a high-alignment part of the configuration plane even though its absolute accuracy is poor: the same multiplicative correction that raises British GNN RMSE by 21% lowers the Australian value by 30%. This is evidence about correction direction at evaluated configurations, not a claim that the Australian pipeline is deployment-accurate or that a rule transfers to an unseen electrical network.

Each region-configuration observation receives $(\rho, \sigma_c/(2\sigma_{\text{res}}))$ from the effective substation factors. Here, *predicted side* denotes the criterion's output, *observed sign* denotes the measured loss change, and *direction accuracy* is their agreement rate. The 435-row evaluation set contains 15 multiplicative configurations per region: three signals on Uniform, three on GPM, and the same three signals under each of three GNN seeds. The nine GNN rows are therefore repeated within-region configurations, not independent units. Additive corrections are outside this evaluation set because they lie outside the criterion's premise. Direction accuracy is 425/435 (97.7%) for log-MSE and 391/435 (89.9%) for original-scale RMSE, compared with a 325/435 (74.7%) majority-class base rate for the latter. Outside the displacement band, original-scale RMSE direction accuracy is 321/331 (97.0%) versus an 83.1% base rate; inside the band it is 70/104 (67.3%) versus 51.9% (Table D.30). The observed signs of log-MSE and original-scale RMSE agree in 391 rows (89.9%). The equal totals in the latter two counts are independently verified: 386 rows are correct in both steps, while five rows contain offsetting errors.

The log-to-linear agreement is 84.6% in Britain, 96.7% in Australia, and 93.3% in Germany. Threshold accuracy for original-scale RMSE are 83.3%, 98.3%, and 93.3%, respectively. The main loss-domain gap occurs on learned bases: the British GNN configurations have 75% log-to-linear sign agreement. This pointwise translation rate does not alter the separate aggregate finding that multiplicative NP correction raises British GNN region-mean RMSE by 21% under the registered paired test; the two values describe different estimands.

The 435 rows are repeated configurations nested in 29 region clusters (16 Britain, 12 Australia, one in Germany), not 435 independent study cases. Region-cluster bootstrap intervals ($B = 2 \times 10^4$) for RMSE-sign classification are 75.4–90.8% in Britain and 96.7–100% in Australia; no interval is interpreted for the single German region. These results support a configuration-level empirical classifier and do not establish cross-case transfer.

A post-hoc robustness probe applies the same threshold to 84 multiplicative configurations on newly constructed bases. It gives 97.6% accuracy for log-MSE and 89.3% for original-scale RMSE. Correction helps in 48.8% of these rows, so the majority-class base rate is 51.2%. This probe reuses the same British and Australian data and correction factors; it is not an independent validation or a test of additive correction. The static-side claim is therefore restricted to the original 96 Uni/GPM configurations. On the British EqualSubregionGrid base, for example, postNP improves RMSE in 0/16 regions.

The three German EqualSubregionGrid configurations form a separate post-hoc robustness probe and are not merged into the 435-row evaluation set. The strict log-MSE direction is correct for all three multiplicative corrections, whereas original-scale RMSE is correct for only one. All three corrections worsen RMSE, so an always-harm rule is 3/3 in this subgroup. The two misclassified margins, about 0.03, lie inside the displacement band. With $n = 3$ in one region, this result identifies a local failure direction but does not estimate its frequency.

Five British regions contain six zero-demand substations. At the default floor, their region-averaged σ_{res} is 1.344, compared with 0.798 elsewhere, while their alignment is lower (0.341 versus 0.497). Removing the zero-demand substations moves nine of the 240 British configurations from the predicted-help side to the predicted-harm side; the global correct counts become 427/435 for log-MSE and 395/435 for original-scale RMSE, compared with 425/435 and 391/435 under the default protocol. Across floors from 10^{-8} to 10^{-4} of regional demand, the six sensitivity conditions span 425–430 correct log-MSE labels and

389–396 correct RMSE labels (Table E.31). The aggregate counts remain within these ranges, but the configuration-level side assignments are not invariant.

The isoline analysis draws 200 alternative allocations for each of the 48 British (region, seed) units from a specified Dirichlet measure whose individual samples lie exactly on the reconstruction-loss isoline. The finite sample is a Monte Carlo reference distribution: it neither exhausts the high-dimensional isoline nor proves that sampled allocations are reachable by the shared GNN parameterization. It is not a null distribution for hypothesis testing.

Within this reference distribution, the learned solution has a median percentile of 0.00% and outperforms all 200 draws in 31 of 48 units; the Uniform-base control has a median percentile of 39%. The median sampled 5th–95th percentile RMSE spread is 0.50, and the median observed sample range is 1.57. These quantities describe the chosen sampling measure and finite draw count; they are not extrema or bounds on every loss-equivalent allocation. The analysis shows that the learned solution lies in a favorable sampled tail while leaving the global identifiability objection open.

5.6. Sensitivity to correction strength

The GNN is trained by self-supervision, without substation-level demand labels. In the deployment settings that this weighting method targets, such labels are generally unavailable, so neither the post-correction parameters (α , γ) nor the prior-loss weight (λ) can be tuned against disaggregation RMSE. All hyperparameters in the main experiments were fixed a priori: $\alpha = 1$ and $\gamma = 2$ follow literature defaults; $\lambda = 0.05$ balances the converged magnitudes of the loss terms (land-use reconstruction ≈ 0.01 ; each prior loss ≈ 0.7 – 0.8). The sweeps below do not seek optimal values; they verify that the pre-specified defaults do not fall in a pathological regime.

The NTL scaling coefficient α , the Proximity decay exponent γ , and a joint scaling factor β are each varied across seven levels; every sweep point is evaluated over all $3 \times 4 = 12$ seed-fold combinations (full results are in Appendix B). Single-factor sweeps show shallow, non-significant optima at low intensity ($\alpha = 0.25$: RMSE 9.19; $\gamma = 1.5$: RMSE 8.65; base 9.27). The combined sweep reaches 8.97 at $\beta = 0.5$, a 3.2% improvement over the 9.27 base, before deteriorating to 11.20 at $\beta = 1.0$ (+20.9%; Fig. 6). Several correction intensities numerically improve RMSE on the GNN base, but none yield a statistically significant improvement.

A sweep over $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$ is conducted on a single fold (seed 42, fold 1, four regions) with retraining at each weight (Table B.16). Within the functional range $\lambda \in [0.01, 0.1]$, RMSE varies by less than one unit. At $\lambda \in \{0.2, 0.5\}$, both columns remain above their respective minima at $\lambda = 0.01$; the RMSE table alone does not identify a convergence failure or a monotonic high-weight effect. The default $\lambda = 0.05$ sits in the functional range and was not selected to minimize RMSE. Across the tested ranges, the adverse British learned-base pattern persists; the sweeps are sensitivity checks rather than a causal isolation of the pattern.

5.7. Concentration sensitivity

A remaining candidate explanation is that the antagonism is an artifact of how concentrated the learned allocation is. To test concentration sensitivity within a specified intervention, grid-point demands are rescaled at inference by a power κ and re-normalized within each sub-region ($\kappa < 1$ flattens the allocation, $\kappa > 1$ sharpens it, $\kappa = 1$ is the unmodified pipeline), and the sweep $\kappa \in \{0.25, 0.5, 1, 2, 4\}$ is evaluated for the uncorrected base, the multiplicative dual-signal correction, and the additive one (full curves in Appendix B.3). This intervention changes the allocation distribution as well as its scalar concentration, so it does not isolate concentration from every other property. Three regularities emerge. First, the base is U-shaped in κ and reaches its minimum at $\kappa = 1$ (RMSE 9.27, against 11.87 flattened to $\kappa = 0.25$ and 27.74 sharpened to $\kappa = 4$): the learned concentration is self-consistent, so “the

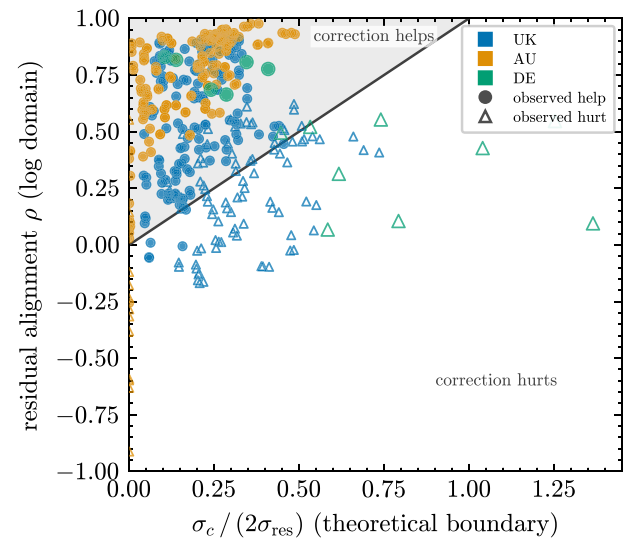


Fig. 5. Configuration-level regime diagram: residual alignment ρ against the strict geometric/log-MSE boundary $\sigma_c/(2\sigma_{res})$ for 435 multiplicative configurations nested in 29 region clusters (Britain 240/16, Australia 180/12, Germany 15/1). Coordinates use the effective substation factor from Section 3.6. The threshold classifies log-MSE direction with 97.7% accuracy and original-scale RMSE direction with 89.9% accuracy, versus a 74.7% majority-class base rate. Original-scale accuracy is 97.0% outside the displacement band and 67.3% inside it; the original-scale result is empirical rather than an if-and-only-if boundary.

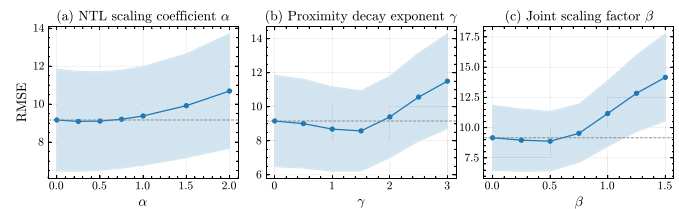


Fig. 6. Post-correction intensity sweep on the GNN base. (a) NTL scaling coefficient α . (b) Proximity decay exponent γ . (c) Joint scaling factor β for the combined NTL + Proximity correction. The dashed line marks the uncorrected GNN base; vertical axes are the 16-region mean RMSE.

allocation is simply too sharp” is not the explanation. Second, the multiplicative correction crosses the regime boundary as κ falls: at $\kappa = 1$ it hurts (11.20; worse in 12 of 16 regions), at $\kappa = 0.25$ it helps everywhere (6.89; better in all 16); a flattened base behaves like the static bases, whose large residuals the factors track, so the boundary crossing of Eq. (13) is observed rather than extrapolated (per- κ values and per-region counts in Table B.17). Third, the additive correction is comparatively insensitive (7.3–9.9 across the sweep, its upper end only at the fully flattened $\kappa = 0.25$), while the multiplicative one swings from 6.9 to 29.9. The power-rescaling intervention is therefore an informative regime coordinate for the multiplicative route in this experiment. The $\kappa = 0.25$ multiplicative pipeline attains 6.89, the lowest single value in the sweep. It is a three-quarters-flattened base plus a multiplicative correction, so most of its information is the correction signal (a uniform base with the same correction, UniNP, reaches 7.45, the flattened GNN adding only 0.56). The most adverse pairing remains a sharp base with a multiplicative correction. In practice this leaves two mitigation strategies: switch the mechanism (additive, tuning-free) or flatten the base and keep the multiplicative correction (requires choosing κ). Under an oracle choice of κ the second is nominally ahead, but its paired difference against the tuning-free additive route (GNNaddP at $\kappa = 1$: 6.89 versus 7.03) is -0.14 (exact two-sided sign-flip permutation over the 16 paired

Table 9

RMSE by region subgroup (load density $\times$ land-use diversity). The n column gives the number of regions per subgroup; subgroups with $n = 1$ report a single region and carry no statistical weight. The table is descriptive and is not used to rank the six reported configurations. Density terciles: low ≤ 0.27 , mid $0.31\text{--}0.40$, high > 0.40 MVA km⁻²; diversity is split at $H_r \approx 1.0$.

Density	Diversity	n	Uni	GPMpostNP	GNN	GNNpostP	GNNpostNP	GNNpriorNP
Low	Low	4	12.56	7.88	9.29	10.41	13.02	9.32
Low	High	1	18.30	6.90	11.94	11.40	13.02	11.68
Mid	Low	2	12.55	6.94	9.09	9.56	11.86	8.66
Mid	High	4	11.19	6.56	7.90	8.23	9.99	7.55
High	Low	2	13.07	5.77	8.16	8.14	9.87	7.44
High	High	3	15.02	8.96	11.04	9.77	10.24	11.04

regions, seeds averaged first, $p = 0.234$; Appendix B.3). The two are statistically indistinguishable, and the additive route avoids an oracle choice of κ . The oracle qualifier marks the practical constraint: selecting the lowest-error κ requires metered substation demand against which candidate values can be evaluated. In the deployment settings targeted by this pipeline, full substation-level labels for tuning are unavailable, though a small metered sample may exist (Section 5.9). The parameter can therefore only be estimated once and frozen, and the sweep is a sensitivity map rather than a tuning recipe.

Training-time concentration is controlled by the softmax temperature τ , the training-side counterpart of κ . Retraining the full pipeline at five initial temperatures spanning a 400-fold range (0.005–2.0; the pipeline default is 0.01) leaves the base nearly unchanged (RMSE 8.83–9.66) and keeps the dual-signal multiplicative antagonism positive at every τ (+0.73 to +2.37; single-seed sweep, five-value table in Appendix B.4). The GNN compensates through its learned edge costs and its trainable temperature: wherever it starts, training re-concentrates the allocation, so the training-time parameter does not cross the regime boundary. The practical observation is limited to the tested intervention: inference-time power rescaling crosses the empirical boundary whereas changing the initial training temperature does not. Choosing the rescaling still requires held-out demand labels and is not proposed as a deployment recipe.

5.8. Regional heterogeneity

To identify conditions under which the synergy–antagonism patterns vary, the 16 regions are cross-stratified by load density ($\delta_r = D_r/\text{Area}_r$; terciles: low ≤ 0.27 , mid $0.31\text{--}0.40$, high > 0.40 MVA km⁻²) and land-use diversity (Shannon entropy H_r ; median split at $H_r \approx 1.0$: low ≤ 1.00 , high ≥ 1.04). Table 9 reports the RMSE for six configurations within each subgroup. With $n = 16$ regions and a 3×2 cross-stratification, subgroup sizes range from 1 to 4 regions; the table should therefore be read as descriptive, flagging potential interaction patterns, rather than as definitive evidence of heterogeneity. The stratification is reported for Britain only: with 12 Australian regions, the same density-tercile $\times$ diversity-median grid would leave about two regions per subgroup, too few even for a descriptive reading.

Three descriptive observations are retained. Percentage differences use $(\text{RMSE}_A - \text{RMSE}_B)/\text{RMSE}_B$, with B as the comparison value. Because each subgroup contains only one to four regions, these observations do not establish a subgroup ranking or a robust ordering of configurations.

First, the difference between GNNpriorNP and GPMpostNP varies across the six subgroups. It is 29% in the high-density, low-diversity subgroup (7.44 versus 5.77, $n = 2$), 18% in the low-density, low-diversity subgroup (9.32 versus 7.88, $n = 4$), and 69% in the low-density, high-diversity subgroup (11.68 versus 6.90). The last value comes from TLC1 alone and carries no subgroup-level statistical weight. The larger differences in the low-density subgroups are consistent with limited information in the five land-use features and with the additional spatial information supplied by the Proximity factor, but the sample is too small to test that explanation.

Second, the high-density, high-diversity subgroup contains London, TLE4, and TLD3. GPMpostNP has RMSE 8.96, while both GNNpriorNP and the uncorrected GNN have 11.04. London dominates the GNN average (16.9, versus approximately 8 for TLE4 and TLD3). After London is removed, the two remaining values are 7.56 for GNNpriorNP and 8.09 for GNN. This change shows that the three-region average is sensitive to London and does not support a general conclusion about high-density, high-diversity regions.

Third, GNNpostNP is close to or above the Uniform RMSE in every subgroup and reaches 13.02 in both low-density subgroups. The observed penalty from the combined multiplicative correction is therefore not confined to one density or diversity group, although the small subgroup sizes prevent us from estimating how the penalty varies with either regional characteristic.

5.9. Internal subsampling diagnostic and always-apply comparison

This experiment tests whether a metered subsample adds useful direction information beyond an always-apply decision. It does not estimate error magnitude or certify planning accuracy. Factor dispersion σ_c is computable from the allocation products and known regional totals, whereas alignment ρ and residual dispersion σ_{res} require metered demand. The diagnostic samples k substations with metered demand, derives F_j through the grid-to-substation relation of Section 3.6, and estimates $\hat{\rho}$ and $\hat{\sigma}_{\text{res}}$ on the subsample. Degenerate subsamples default conservatively to the detrimental side. We evaluate 500 resamples per (region, configuration) pair for $k \in \{3, 5, 8, 10, 15, 20\}$ against the frozen full-sample direction labels. Fig. 7 shows the accuracy curves. At $k = 8$, the British static curve reaches 91.4%, below its 100% always-apply rate. The British GNN curve reaches 62.2%, compared with 38.2% for always applying and 61.8% for the majority-class rule. The full-sample criterion reaches 72.2%, 10.4 percentage points above the majority-class base rate; the shortfall is therefore in the $k = 8$ estimate, not the criterion itself. On eligible Australian rows, the all/static/GNN curves reach 92.2%/94.0%/91.0%, below the corresponding 99.0%/100%/98.4% always-apply rates. These are internal resampling results: the subsamples and full-sample labels come from the same evaluated configuration families, and five Australian regions are ineligible at $k = 8$. Prospective accuracy and false-positive rates on a held-out network remain untested. A separate logistic classifier uses factor dispersion σ_c , base type, and allocation entropy. Trained on the remaining regions, it requires no substation demand labels from the held-out region as prediction inputs. Under leave-one-region-out validation over the 420 British and Australian pairs, it reaches 78.6% accuracy, compared with 76.0% for always applying the correction. The held-out region's country remains represented in training.

5.10. The German Börde case: descriptive sensitivity under a shared synthetic topology

The German Börde district contains one descriptive study case with 13 substations and 34 municipalities. Its GNN base has predictive $R^2 = 0.956$, whereas the squared mean Pearson correlation is 0.963; the

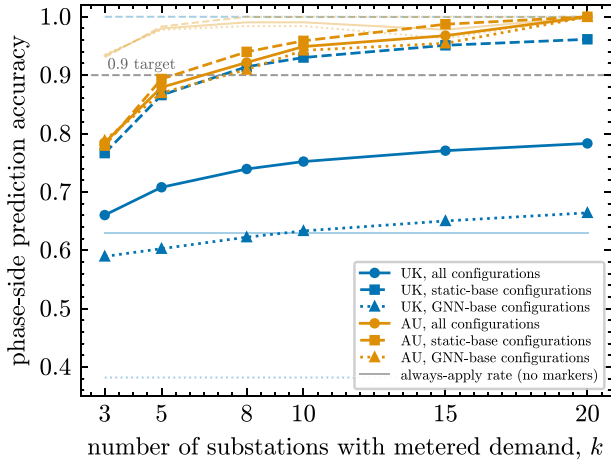


Fig. 7. Internal direction classification from a k -substation metered sample (500 resamples per configuration). Marker curves are the sampled diagnostic; matching pale curves without markers are always-apply accuracies for the same eligible configurations, and the gray dashed line marks 0.9. The sampled coordinates use the effective factors of Section 3.6. The diagnostic falls below always apply on both static-base curves and on all eligible Australian curves.

former is the relevant predictive-accuracy statistic, and neither number turns the case into an independent study-case validation. We report configuration directions and sensitivity under the common synthetic topology without statistical inference or a transfer claim. A jackknife influence check removes one substation at a time from the metric calculation while holding the trained predictions fixed. The three-seed mean LOO correlation remains at least 0.956, with the largest drop at Zielitz (45 MW; cross-seed minimum $r = 0.953$; Appendix C). This shows metric sensitivity to individual nodes; it does not test overfitting.

The configurations show the following within-case differences. EqualSubregionGrid Base has a predictive $R^2 = 0.710$ in this case, and applying any of its six corrections lowers that value. Fifteen existing configurations exceed Grid Base, but none of the seven Grid configurations reaches UniNP at 0.967. These comparisons are descriptive and do not support a cross-case configuration ranking. The German evaluation target is a provider's current-load snapshot reported in 2.5 MW increments, not an interval-metered annual peak series. The uncorrected GNN RMSE of about 2.25 MW is therefore close to one reporting step, which limits the interpretation of predictive $R^2 = 0.956$ and the small correction margins.

UniNP (1.95) cuts RMSE by 70% relative to Uni (6.50), mirroring the strong static-side synergy observed in Britain, and edges out the uncorrected GNN (2.25) by 13%. The static corrections also reduce error in this German configuration set, but one region cannot establish a cross-case regularity.

In Britain, GPMpostNP (7.31 MVA) records a lower RMSE than UniNP (7.45 MVA); in Börde, UniNP (1.95 MW) records a lower RMSE than GPMpostNP (2.33 MW). This reversal is consistent with the single-sample nature of the German evaluation and with the coarse spatial resolution of the economic output data that supplies the GPM weighting signal. This case cannot determine whether the difference is systematic or case-specific.

Post-correction antagonism is larger in the German case: NTL post-correction worsens the GNN base from 2.25 to 3.68 (+64%), compared with +2% in Britain; the combined NTL+Proximity post-correction inflates RMSE to 6.58 (+193%), compared with +21% in Britain (per-configuration values are in Table C.20). In this small case, the GNN has high predictive accuracy ($R^2 = 0.956$), and every tested multiplicative perturbation increases RMSE. The coexistence of these two observations is descriptive; it is not interpreted as a quantified residual-capacity mechanism. The additive mechanism is less adverse in this case

Table 10

Descriptive configuration performance in the German Börde district (RMSE in MW; three-seed mean $\pm$ standard deviation for GNN configurations). The table compares configurations within this study case; it does not test cross-case transfer.

Configuration	RMSE	Corr
Uni	6.50	0.860
GPM	6.88	0.846
UniNP	1.95	0.984
GPMpostNP	2.33	0.977
GNN	2.25 ± 0.05	0.9812 ± 0.0004
GNNaddP	2.42 ± 0.21	0.9756 ± 0.0042
GNNaddNP	3.79 ± 0.33	0.9444 ± 0.0092
GNNpriorNP	2.75 ± 0.05	0.9710 ± 0.0012

Table 11

Prior-loss RMSE changes relative to the unmodified GNN on Britain, Australia, and Germany. No tested prior configuration produces a validated significant benefit, and the table is not used to define a mechanistic applicability rule. The Australian NTL+Proximity change is -0.003 and rounds to -0.00 .

Prior config.	Britain Δ RMSE	Australia Δ RMSE	Germany Δ RMSE
NTL only	-0.08 (-0.9%)	$+1.01$ ($+2.5\%$)	$+0.51$ ($+22.7\%$)
Prox only	-0.12 (-1.3%)	-0.04 (-0.1%)	$+0.34$ ($+15.3\%$)
NTL + Prox	-0.24 (-2.6%)	-0.00 (-0.0%)	$+0.50$ ($+22.4\%$)

(three-seed means, Table 10; the multiplicative counterparts appear in Table C.20): GNNaddP reaches 2.42 against 4.67 for GNNpostP (-48%), and GNNaddNP reaches 3.79 against 6.58 for GNNpostNP (-42%). These are within-case correction-procedure comparisons.

The Grid construction also exposes the role of Voronoi crossing. Positive-demand municipalities contribute 33.3% of their mass through substation cells whose substations lie outside the source sub-region; all 13 substation cells intersect more than one municipality. The ratio of positive-demand source sub-regions to substations is $11/13 = 0.846$, not a one-to-one mapping. These facts help explain why the grid-level reference and direct substation-level EqualSubregion answer different questions.

A jackknife leave-one-out analysis across configurations shows that multiplicative correction degrades robustness to single-point removal alongside average accuracy. The standard deviation of LOO correlation rises from 0.008 (uncorrected GNN) through 0.030 and 0.038 (single post-corrections) to 0.051 (combined NTL+Proximity); the minimum LOO correlation falls from 0.953 to 0.690. Removing one substation from the combined post-correction can reduce correlation by 0.17 points, showing sensitivity to individual nodes within this descriptive case.

Prior-loss integration is likewise adverse in the German case: dual prior-loss training raises GNN RMSE from 2.25 to 2.75 (+22%). The per-component breakdown (Table 11) shows that all three prior configurations degrade performance.

The predictive- R^2 audit changes the interpretation without changing the observed direction of the prior-loss results. The German GNN has predictive $R^2 = 0.956$; GNNpriorN, GNNpriorP, and GNNpriorNP decrease it to 0.934, 0.941, and 0.935. Britain and Australia likewise show no validated significant prior-loss benefit. Squared correlation is not predictive R^2 , so the former correlation-derived fraction and capacity interpretation are withdrawn. The evidence supports only the null-result statement: no tested prior-loss configuration yields a significant benefit, and this study does not establish a mechanism that predicts when prior-loss integration will help.

The NTL-Proximity correlation in the German case is 0.258, but with $N = 13$ substations, a correlation test has negligible statistical power; the redundancy-amplification hypothesis cannot be confirmed or refuted here.

Table 12

Relative power-flow sensitivity to alternative demand inputs under the shared synthetic Börde 110 kV MST and centroid slack bus (configurations ordered by RMSE ascending; values do not represent the actual Avacon network). Δ_{MAE} : per-line mean absolute deviation of line loading from the true-load reference (pp). ℓ_{max} : absolute maximum line loading (%). EqualStation assigns each substation an equal share of the total district demand.

Configuration	RMSE (MW)	Δ_{MAE} (pp)	ℓ_{max} (%)
True load	— — —	0.00	132.9
UniNP	1.95	0.74	132.1
GNN	2.25	1.27	129.3
GNNpriorNP	2.75	1.55	129.5
EqualStation	10.77	17.03	102.5

Downstream sensitivity is measured by comparing power-flow outputs in Table 12 after each demand estimate is injected into the same simplified 110 kV model (Section 4.2). The per-line loading deviation Δ_{MAE} ranges from 0.74 pp for UniNP to 17.03 pp for EqualStation, which assigns an equal share of total district demand to every substation. In this radial MST with one remote high-demand node, maximum loading is 132.1% for UniNP and 102.5% for EqualStation. These values show relative input sensitivity under fixed synthetic assumptions, not loading estimates for the actual Avacon network or evidence to support a planning decision.

6. Discussion

The effect of auxiliary information depends on both its integration mechanism and its base allocation.

6.1. Mechanistic interpretation

On the British learned base, multiplicative correction at unit intensity increases RMSE whereas the specified additive procedure decreases it. This contrast changes both the functional form and the intensity because the additive scale α_s adapts to sub-region dispersion while the multiplicative arm uses a unit exponent. The intensity sweep confirms the distinction: $\beta = 0.5$ lowers RMSE from 9.27 to 8.97, whereas $\beta = 1$ raises it to 11.20. The controls exclude a simple double-normalization explanation but do not isolate multiplication as the sole cause.

Voronoi aggregation does not preserve the products of separately aggregated factors, so the observed factor association and post-aggregation RMSE interaction remain empirical. Together, the theory and experiments support re-evaluating an inherited correction when a learned weighting method replaces a rule-based one.

Prior-loss integration produced no validated RMSE benefit in any study case, and the experiments do not identify when it should help. Some post-corrections also increase correlation while increasing RMSE. Correlation describes spatial co-variation, whereas RMSE measures absolute error; reporting both prevents a correlation gain from masking poorer demand estimates.

6.2. Engineering significance and accuracy boundaries

The British change from Uni (13.10 MVA) to GPMpostNP (7.31 MVA) shows that integration materially changes a planning input. It does not imply an equivalent planning-margin reduction, which would require a reliability criterion, an error distribution, and a capacity-decision rule. The German synthetic-network analysis establishes sensitivity to alternative demand inputs.

Nighttime Light and Proximity improve the tested uniform-weighting and GPM bases without training, but the results do not extend to every static construction.

The auxiliary layers can be incorporated into Voronoi-based modules such as those used by PyPSA-Eur [1]. The form of their integration and

absolute accuracy must still be validated in the target pipeline before quantitative planning use.

The evidence here concerns the direction of the correction effect, not the absolute accuracy of the resulting demand estimates; on that second question the results are unfavorable under both reported estimands. Region-mean predictive R^2 is negative for all 22 configurations in both main study cases. Selected British pooled values are positive, but the direct EqualSubregion attains 0.248 in Britain and 0.194 in Australia without the grid/Voronoi route. Moreover, London and Sydney City and Inner South contribute 47.3% and 52.2% of the pooled denominators. These results do not support deployment-level accuracy claims based on the relative correction effects.

6.3. Limitations

The evidence comprises 16 British regions, 12 Australian regions, and one descriptive German region. The 435 configuration rows are nested within 29 region clusters and are not independent study-case validations. Other settlement, electrification, or generation patterns require re-estimation of the criterion.

The fixed GNN architecture and self-supervised objective limit generalization across model classes. Better features or supervised demand labels could change the learned base and its interaction with auxiliary information. Nighttime Light may also lose discrimination in low-electrification settings, while Proximity can be misleading where substations primarily serve generation.

The aggregate direction result is stable across the tested numerical-floor conditions, but some configuration-level side assignments change when zero-demand substations are excluded.

The German target is a provider's current-load snapshot reported in 2.5 MW increments, so a base RMSE of about 2.25 MW sits close to the reporting step and the predictive $R^2 = 0.956$ is descriptive. The original-scale RMSE direction is correct for only one of the three German grid corrections; the two missed margins, of about 0.03, fall inside the displacement band.

Finally, the study evaluates annual peak demand and only considers multiplicative and additive post-correction. Temporal profiles and alternative correction functions remain untested. The theorem covers positive multiplicative factors under geometric re-normalization and log-MSE; the additive procedure and original-scale RMSE results are empirical comparisons rather than derived guarantees.

7. Conclusion

This study crosses the weighting method and the integration mechanism under fixed Voronoi assignment to examine substation-level load disaggregation. The main evidence comprises metered peak demand from 1891 British and 143 Australian substations; the 13-substation German case is descriptive.

Under geometric re-normalization and log-MSE, the conditional-benefit threshold is exact. In the implemented pipeline, it classifies the original-scale RMSE direction in 391/435 configurations (89.9% versus a 74.7% majority-class base rate); accuracy is 97.0% outside the displacement band and 67.3% inside it. The result depends on both base allocation and integration mechanism. At unit multiplicative intensity ($\beta = 1$), correction cuts RMSE by 37–43% on the tested uniform and GPM bases, raises the British GNN value by 21%, and lowers the Australian one by 30%. Absolute accuracy remains study-case- and estimand-dependent, so these relative effects do not establish deployment-ready demand estimates.

The grid-to-substation mapping matches stored correction factors up to machine precision and separates the exact geometric/log-MSE result from the empirical arithmetic/RMSE classifier. The contribution is therefore directional: given metered demand, the criterion classifies whether a multiplicative auxiliary correction helps or harms an evaluated base allocation, nor does it calibrate the resulting error magnitude. Its value as a pre-deployment decision rule is not established

here: subsample estimates show no consistent advantage over simple decision rules (Section 5.9). The findings inform integration-mechanism choice. Future work should derive conditions for other correction functions, develop and independently validate estimators of factor-residual alignment and residual dispersion using historical or partial observations, test independent electrical networks, and extend the analysis to temporal demand profiles.

CRedit authorship contribution statement

Xuanhao Mu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kundan Thota:** Writing – review & editing, Methodology. **Nan Liu:** Writing – review & editing, Methodology. **Jianlei Liu:** Writing – review & editing, Writing – original draft, Visualization. **Thorsten Schlachter:** Writing – review & editing, Supervision, Investigation. **Veit Hagenmeyer:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Gemini (Google DeepMind), ChatGPT (OpenAI) and Claude (Anthropic) in order to perform text polishing and code debugging for plotting. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported in part by the Helmholtz Association through the project “Helmholtz Platform for the Design of Robust Energy Systems and Their Supply Chains” (RESUR I). Its successor project RESUR II (grant number 03EI1123 A) has been funded by the German Federal Ministry for Economic Affairs and Energy. The funding was provided within the scope of energy research, based on a decision by the German Bundestag. Additional support has been provided by the Helmholtz AI HAICORE partition (HAICORE@KIT).

Appendix A. GNN architecture and training details

This appendix provides a self-contained description of the HGNN used as the GNN demand weighting method throughout this paper. Full details are available in the companion paper [11]; we reproduce the essential elements here to ensure reproducibility.

A.1. Graph construction

Each ITL-2 region is represented as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with two node types. **Sub-region nodes** ($\mathcal{V}_s$) correspond one-to-one with the ITL-3 sub-regions within the ITL-2 zone; each carries a 5-dimensional feature vector $\mathbf{x}_s = (\text{pct}_s^{\text{res}}, \text{pct}_s^{\text{com}}, \text{pct}_s^{\text{ind}}, \text{pct}_s^{\text{agr}}, \text{pct}_s^{\text{oth}})$ representing the regional electricity consumption structure by land-use class, together with a scalar demand attribute D_s . Their implementation node type is **source**. **Grid-point nodes** ($\mathcal{V}_g$) are $\sim 50,000$ spatial sample points per region, each carrying the same 5-dimensional land-use proportion vector

$\mathbf{x}_a = (p_a^{\text{res}}, \dots, p_a^{\text{oth}})$ as described in Section 4.1. Their implementation node type is **agent**.

Two directed edge types enable bidirectional message passing: (source, connects_to, agent) links each sub-region node to all grid-point nodes within its ITL-3 boundary, and (agent, rev_connects_to, source) provides the reverse direction.

A.2. Encoder architecture

The encoder maps heterogeneous node features to a shared embedding space through a three-stage pipeline. A per-node-type linear layer first projects input features to the hidden dimension:

$$\mathbf{h}_v^{(0)} = \mathbf{W}_{\text{type}(v)} \mathbf{x}_v + \mathbf{b}_{\text{type}(v)}, \quad \mathbf{W} \in \mathbb{R}^{d_h \times d_{in}} \quad (\text{A.1})$$

The hidden dimension is $d_h = 256$. Baseline and prior-loss configurations use five land-use features for both source and grid-point nodes. In GNNfusion, the source input remains five-dimensional, while each grid-point input is seven-dimensional after appending region-wise standardized $\log(1 + \text{NTL})$ and $\log(1 + \text{Proximity})$. The type-specific input projections are therefore $\mathbf{W}_{\text{source}} \in \mathbb{R}^{256 \times 5}$ and $\mathbf{W}_{\text{grid}} \in \mathbb{R}^{256 \times 7}$ for the fusion configuration; all subsequent HGT layers and the 128-dimensional output head are unchanged. These initial representations are then refined through three stacked Heterogeneous Graph Transformer (HGT) layers [32] with multi-head attention ($H = 4$ heads), per-layer residual connections, layer normalization, and dropout ($p = 0.1$):

$$\mathbf{h}_v^{(\ell+1)} = \text{LayerNorm}\left(\mathbf{h}_v^{(\ell)} + \text{Dropout}\left(\text{HGT}^{(\ell)}\left(\{\mathbf{h}_u^{(\ell)}\}_{u \in \mathcal{N}(v)}\right)\right)\right) \quad (\text{A.2})$$

for $\ell = 0, 1, 2$. Finally, a linear output head maps the hidden representations to the embedding dimension $d_e = 128$, followed by L2-normalization and a learnable *per-dimension* scaling vector $\mathbf{g}_v \in \mathbb{R}^{d_e}$:

$$\mathbf{z}_v = \mathbf{g}_v \odot \frac{\mathbf{W}_{\text{out}} \mathbf{h}_v^{(L)}}{\|\mathbf{W}_{\text{out}} \mathbf{h}_v^{(L)}\|_2} \quad (\text{A.3})$$

where $\odot$ denotes element-wise multiplication.

A.3. Edge weight prediction

Allocation weights are derived from the learned embeddings of source-grid-point pairs in two steps. For each edge (s, a) , a cost score combines embedding distance with a learned gating function:

$$c_{sa} = \|\mathbf{z}_s - \mathbf{z}_a\|_2 \cdot \text{sigm}(\text{MLP}(\|\mathbf{z}_s\|_2)) \quad (\text{A.4})$$

where sigm denotes the logistic sigmoid function and the MLP has architecture $[2d_e \rightarrow d_e \rightarrow 1]$ with ReLU activation.

These costs are then normalized per source via a softmax with learnable temperature τ :

$$w_{sa} = \frac{\exp(-c_{sa}/\tau)}{\sum_{a' \in \mathcal{N}(s)} \exp(-c_{sa'}/\tau)} \quad (\text{A.5})$$

as stated in Eq. (6). The temperature is parameterized as $\tau = \exp(\log \tau_0)$ with $\log \tau_0$ initialized to $\log(0.01)$.

A.4. Training procedure

All hyperparameters are summarized in Table A.13; this section details the non-obvious design choices.

Table A.13
GNN hyperparameters used in all experiments.

Parameter	Value
<i>Encoder</i>	
Hidden dimension d_h	256
Embedding dimension d_e	128
Number of HGT layers	3
Attention heads	4
<i>Edge weighting</i>	
Initial temperature τ_0	0.01
Gate MLP	[256 $\rightarrow$ 128 $\rightarrow$ 1]
<i>Optimization</i>	
Optimizer	AdamW
Learning rate η	10^{-3}
Weight decay	10^{-4}
Epochs	200
<i>LR schedule</i>	
Warmup	20 epochs, $0.1\eta \rightarrow \eta$
Cosine annealing	to $\eta_{\min} = 10^{-5}$
Final decay	20 epochs
<i>Prior loss</i>	
λ_{ntl}	0.05
λ_{prox}	0.05
<i>Regularization</i>	
Dropout	0.1

Loss and optimization

Training minimizes the self-supervised land-use prediction loss (Eq. 7). For prior-loss configurations, the total objective is augmented as in Eq. (23) with fixed scalar weights $\lambda_{\text{ntl}} = \lambda_{\text{prox}} = 0.05$. The optimizer is AdamW with learning rate $\eta = 10^{-3}$ and weight decay of 10^{-4} .

A three-phase learning rate schedule is used:

1. *Warmup* (epochs 1–20): linear ramp from 0.1η to η .
2. *Cosine annealing* (epochs 21– $E-20$): decay from η to $\eta_{\min} = 10^{-5}$.
3. *Linear decay* (final 20 epochs): from $\eta_{\min}$ to 0.01η .

where $E = 200$ epochs for single-prior and base configurations, and $E = 400$ for the dual-prior (NTL + Prox) configuration.

Multi-graph training and reproducibility

Each epoch iterates over all training-set region graphs (12 regions in the K-fold). Model selection uses the epoch with the lowest aggregate training loss across all graphs. Three random seeds (42, 123, 456)

control both parameter initialization and K-fold region assignment. All experiments use a single NVIDIA 4070 Ti Super; mixed precision is not employed.

A.5. Loss implementation details

Land-use reconstruction loss

The target $\mathbf{q}_s^{\text{true}}$ of Eq. (7) is the vector of five class-percentage entries of sub-region s , normalized to sum to one. The aggregation matrix $\mathbf{M}$ is one-hot: during graph construction each grid point is assigned its dominant land-use class $k^*(a)$ (the argmax of its 5-dimensional land-use proportion vector), so each row of $\mathbf{M}$ contains exactly one nonzero entry; this structure has been verified programmatically for every edge of all 16 region graphs. The reconstructed mixture is therefore the vector of class masses $\sum_{a \in \mathcal{A}_s: k^*(a)=c} w_{sa}$, normalized across classes. Both the target and the reconstructed shares are floored at 10^{-8} before the logarithm; no other smoothing is applied.

Prior losses

The prior target of Eq. (21) is computed in the following order: the auxiliary value $v(a)$ is set to zero for grid points outside RCI, then transformed by $\log(1 + \cdot)$, floored at 10^{-8} , and normalized per source to give $q_{s,a}$; inside the KL term of Eq. (22), both $q_{s,a}$ and w_{sa} are floored at 10^{-8} . An off-support target thus equals the floor divided by the per-source normalizer rather than a data-driven small value, and the exact-zero support stated in Section 3.8 holds up to the floor. We measured the actual gradient share of these floor-level terms on the trained dual-prior models (three seeds, 255 sources): in the $\partial L / \partial w$ decomposition, the off-support share has median 1.9% per source but is heavy-tailed (90th percentile 23%; up to 77% for rural sources where the RCI mask zeroes most grid points), so off-support terms are not uniformly negligible. Two properties bound their effect: under the softmax parameterization the off-support pull is $w_{sa} - q_{s,a} \approx w_{sa} > 0$ in logit space, i.e., it pushes off-support weights toward the floor (the direction the prior intends) rather than injecting a data-driven signal; and the double-sided floor gates the gradient of every edge with w_{sa} below the floor to exactly zero (about 8.6×10^4 edges per seed in the measured pool). Sources whose masked auxiliary mass $\sum_{a \in \mathcal{A}_s} \log(1 + v(a))$ falls below 10^{-6} (no RCI grid point with a nonzero auxiliary signal) are excluded from the per-source average of Eq. (22).

A.6. Input-channel disclosure

Table A.14 itemizes every channel through which information reaches the GNN. All entries were generated programmatically (by reading configuration objects from the training code, statically scanning its source, and measuring the cached graphs used in all experiments) rather than being transcribed by hand.

Table A.14

Input-channel disclosure for the GNN. Entries are machine-generated from the training code and the cached region graphs; “measured” denotes properties verified per graph on the caches used in all experiments.

Channel	Disclosure	Verified from
Grid-point node features (baseline/prior)	Five land-use proportion columns only; contain no NTL and no coordinates	training script (column list read programmatically)
Grid-point node features (GNNfusion)	Seven columns: the five land-use proportions plus region-wise z-scored $\log(1 + \text{NTL})$ and $\log(1 + \text{Proximity})$; geometry remains excluded	‘027_train_feature_fusion.py’ column list and graph builder
Direct grid-point connectivity	None configured in the main pipeline	training script constant
Training objectives	Baseline: land-use reconstruction loss only; NTL/Proximity: one prior term added ($\lambda = 0.05$); NTL + Prox: both priors ($\lambda = 0.05$ each); 200 epochs (400 for the dual-prior configuration)	configuration map (read programmatically)
Initial softmax temperature	$\tau_0 = 0.01$ at both occurrences in the script	static source scan
Geometry handling	Geometry column dropped in the first feature-preprocessing step	graph-construction module
Coordinates	Stored as a separate graph attribute; never part of the node feature matrix	graph-construction module
Encoder forward pass	Receives node features and edge indices only; coordinates, NTL, Proximity, and the RCI mask do not enter	training and inference call sites
NTL pathway	Absent from node features; affects training only through the prior loss of the NTL and NTL + Prox configurations; unused at inference	loss registry + configuration map
Proximity pathway	Absent from node features; affects training only through the prior loss of the Proximity and NTL + Prox configurations; unused at inference	loss registry + configuration map
Main graph caches	16 region graphs; baseline/prior feature dimension 5 with columns matching the declared list; star membership structure; zero grid-point–grid-point edges; 821,182 grid points in total	cached graphs (measured)
Fusion graph caches	Separate cache family; grid-point feature dimension 7 and source feature dimension 5; identical region/fold/seed protocol to the baseline	fusion cache manifest and training call sites
Adjacency control variant	Identical node features; 3,349,214 explicit grid-point–grid-point edges added, confirming that the zero count above is measured rather than definitional	cached control graphs (measured)

Appendix B. Hyperparameter robustness check

This appendix reports the numerical results behind the sensitivity analysis in Section 5.6.

B.1. Post-correction intensity sweep

Post-correction is applied post-hoc to a pre-trained GNN base; no retraining is required. Strength is parameterized as follows: for NTL, the factor is interpolated toward neutrality, $f_\alpha(a) = 1 + \alpha(f(a) - 1)$, so that $\alpha = 0$ is the uncorrected base and $\alpha = 1$ is the pipeline factor; for Proximity, the decay exponent γ is varied directly (default $\gamma_0 = 2$); the joint factor β applies both simultaneously, scaling the NTL factor with $\alpha = \beta$ and setting the Proximity exponent to $\beta\gamma_0$. Each sweep point is evaluated across all $3 \times 4 = 12$ seed–fold combinations. Table B.15 reports the 16-region mean RMSE at each intensity level.

Table B.15

Post-correction intensity sweep on the GNN base (16-region mean RMSE, 12 seed–fold evaluations per point). α : NTL scaling coefficient; γ : Proximity decay exponent; β : joint NTL + Proximity scaling factor. An intensity of zero corresponds to the uncorrected GNN base.

NTL (α)		Proximity (γ)		Combined (β)	
α	RMSE	γ	RMSE	β	RMSE
0.00	9.27	0.0	9.27	0.00	9.27
0.25	9.19	0.5	9.12	0.25	9.07
0.50	9.20	1.0	8.79	0.50	8.97
0.75	9.29	1.5	8.65	0.75	9.58
1.00	9.45	2.0	9.42	1.00	11.20
1.50	10.00	2.5	10.57	1.25	12.89
2.00	10.81	3.0	11.49	1.50	14.22

For NTL, the minimum RMSE (9.19) at $\alpha = 0.25$ is a non-significant 0.9% improvement over the base; RMSE increases monotonically beyond this point. For Proximity, the minimum (8.65) at $\gamma = 1.5$ is a 6.7% improvement, but this optimum does not transfer to the combined correction: the default $\gamma = 2.0$ already produces antagonism when NTL is also applied. For the combined sweep, any $\beta > 0.5$ produces an RMSE worse than the base.

B.2. Prior-loss weight sweep

The prior-loss sweep requires retraining at each weight setting. To limit computational cost, a single fold (seed 42, fold 1) covering four regions is used. Table B.16 reports the mean RMSE at each λ .

Table B.16

Prior-loss weight sweep (seed 42, fold 1, 4 regions). RMSE is the mean across the four test regions.

λ	NTL prior RMSE	Proximity prior RMSE
0.01	10.48	10.30
0.05	10.55	11.03
0.10	10.50	10.84
0.20	11.09	10.51
0.50	10.78	10.69

The Proximity prior shows a wider gap between $\lambda = 0.01$ (10.30) and the default $\lambda = 0.05$ (11.03) than NTL does (10.48 vs 10.55). This gap is dominated by a single fold that includes London (GNN RMSE ≈ 16.9), the outlier identified in Section 5.8. At $\lambda \in \{0.2, 0.5\}$, both columns remain above their respective minima at $\lambda = 0.01$. Because only RMSE is recorded in this sweep and the Proximity column is non-monotonic, the table does not establish that the main loss is overpowered or that convergence degrades at $\lambda \geq 0.2$.

B.3. Inference-time concentration rescaling

Fig. B.8 reports the full concentration sweep in Section 5.7: grid-point demands are raised to the power κ and re-normalized within each sub-region before Voronoi aggregation, for the uncorrected GNN base, the multiplicative dual-signal correction, and the additive one.

Table B.17 gives the numerical grid behind the curves, together with the number of regions in which the multiplicative correction is worse than the base at each κ . The oracle comparison of Section 5.7 pairs the $\kappa = 0.25$ multiplicative pipeline (6.89) with GNNaddP at $\kappa = 1$ (7.03, Table 3): the mean paired difference over the 16 regions is -0.14 (seeds averaged first, exact two-sided 2^{16} sign-flip permutation, $p = 0.234$; frozen close-out artifact of the temperature-sweep experiment).

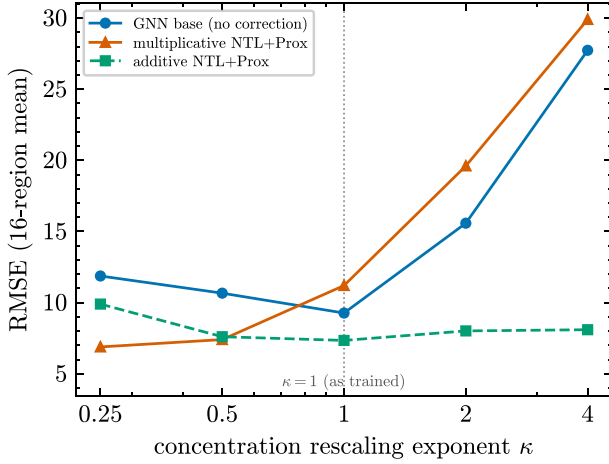


Fig. B.8. Concentration rescaling sweep (British case, base configuration; log-scaled κ axis). The uncorrected base is U-shaped with its optimum at the unmodified $\kappa = 1$; the multiplicative dual-signal correction crosses from harmful to helpful as κ falls; the additive correction is nearly flat. κ raises grid-point demands to a power and re-normalizes within each sub-region: $\kappa < 1$ flattens, $\kappa > 1$ sharpens, $\kappa = 1$ is the unmodified pipeline.

Table B.17

Concentration rescaling grid (British case, base configuration): 16-region mean RMSE of the uncorrected base, the multiplicative dual-signal correction, and the additive one at each κ , and the number of regions (of 16) in which the multiplicative correction is worse than the base.

κ	Base	Mult. NP	Add. NP	Regions worse
0.25	11.87	6.89	9.90	0
0.5	10.67	7.40	7.60	2
1	9.27	11.20	7.34	12
2	15.58	19.61	8.01	15
4	27.74	29.90	8.10	11

B.4. Training-time temperature sweep

Table B.18 reports the training-time sweep in [Section 5.7](#): the full pipeline is retrained at five initial softmax temperatures (single seed; the default-temperature run lies within the cross-seed interval of the main experiments, $9.30 \in [9.15, 9.49]$, where the interval endpoints are the per-seed 16-region mean RMSE values of the base configuration, 9.16/9.49/9.15 for seeds 42/123/456). The base varies by less than one RMSE unit over a 400-fold temperature range, and the dual-signal multiplicative antagonism stays positive throughout.

Table B.18

Training-time temperature sweep (British case, seed 42): base and multiplicative dual-signal RMSE, and the antagonism Δ between them, for five initial temperatures.

Initial τ	Base RMSE	Mult. NP RMSE	Δ
0.005	9.66	11.34	+1.67
0.01	9.30	11.26	+1.96
0.05	8.86	11.24	+2.37
0.5	8.83	10.24	+1.42
2.0	9.13	9.86	+0.73

Appendix C. Jackknife leave-one-out influence analysis (German Börde)

This appendix reports the full jackknife leave-one-out (LOO) analysis for the German Börde case discussed in [Section 5.10](#). Each of the 13 substations is removed in turn; Corr and RMSE are recomputed for the remaining 12. Results are averaged across three random seeds (42, 123, 456). Predictions are held fixed and the model is not retrained. The analysis measures influence on the reported metrics, not train-test overfitting.

Consistent with the protocol of [Section 4.3](#), no statistical inference is attached (13 substations); the tables provide a qualitative stability diagnostic.

C.1. Per-substation LOO stability

Table C.19 reports the per-substation LOO results for the base GNN configuration. The “Predicted” column gives the three-seed mean of the GNN-allocated demand; “ $|\Delta|$ ” is the mean absolute error for the dropped substation. LOO Corr and LOO RMSE are computed over the 12 remaining substations after removal.

Table C.19

Per-substation jackknife LOO results for the base GNN configuration (three-seed mean). Substations are sorted by the provider’s load target. Oswl is the provider identifier for the record whose name field is empty. Removing any single substation leaves LOO Corr ≥ 0.956 ; no single point drives the high full-sample correlation ($r = 0.981$).

Substation	Target (MW)	Pred. (MW)	$ \Delta $ (MW)	Corr	RMSE (MW)
Wellen	2.5	3.3	0.79	0.980	2.32
Badeleben	5.0	5.7	0.67	0.980	2.33
Erleben	5.0	6.2	1.21	0.981	2.31
Weferlingen	7.5	7.7	0.20	0.980	2.34
Haldensleben	10.0	12.5	2.48	0.984	2.22
Santersleben	15.0	10.0	5.00	0.989	1.83
Wolmirstedt	15.0	12.1	2.90	0.984	2.18
Barleben	17.5	17.1	0.42	0.981	2.33
Flechtingen	17.5	17.5	0.14	0.981	2.34
Oswl	22.5	21.9	2.15	0.982	2.25
Oschersleben	22.5	21.8	0.67	0.981	2.33
Osterweddingen	22.5	23.1	2.16	0.982	2.25
Zielitz	45.0	48.6	3.60	0.956	2.09
Full sample	—	—	1.72	0.981	2.25

Zielitz (45 MW) is the most influential substation: its removal produces the largest drop in LOO Corr (from 0.981 to 0.956 on average; minimum across seeds: $r = 0.953$, seed 123). This substation carries the highest demand in the district, more than twice the next largest, so its removal disproportionately reduces the dynamic range available for correlation estimation. Even so, LOO Corr = 0.953 remains high; the base performance reflects genuine spatial structure rather than dependence on a single outlier.

Santersleben (15 MW) has the largest absolute prediction error ($|\Delta| = 5.00$ MW), yet its removal increases LOO Corr to 0.989; this substation is the GNN’s most poorly predicted point. The pattern is consistent: substations whose removal improves Corr are those for which the GNN prediction deviates most from the regional trend.

C.2. LOO stability across configurations

Table C.20 compares the LOO stability of all four GNN-based configurations under the base configuration. The gap between full-sample

Table C.20

LOO stability comparison across GNN configurations (base configuration, three-seed aggregation over $13 \times 3 = 39$ LOO iterations per configuration). ΔCorr = LOO mean – full-sample mean; values near zero indicate limited single-node influence, not absence of overfitting.

Configuration	Full Corr	LOO Corr	LOO min	Full RMSE	LOO RMSE
GNN	0.981	0.980 ± 0.008	0.953	2.25	2.24 ± 0.15
GNNpostN	0.950	0.947 ± 0.030	0.819	3.68	3.67 ± 0.73
GNNpostP	0.923	0.918 ± 0.038	0.770	4.67	4.65 ± 0.74
GNNpostNP	0.863	0.858 ± 0.051	0.690	6.58	6.54 ± 0.81

and LOO metrics is small for the uncorrected GNN, indicating limited influence of any single measured node on the reported metrics. Post-correction configurations show wider LOO variability; the combined multiplicative correction (GNNpostNP) has the largest spread.

Two patterns are noteworthy. First, the LOO variability (standard deviation of LOO Corr) increases monotonically from GNN (0.008) through single post-corrections (0.030–0.038) to the combined correction (0.051). This reflects the compounding sensitivity of multiplicative corrections to individual data points under conservation constraints. Second, the minimum LOO Corr degrades from 0.953 (GNN) to 0.690 (GNNpostNP), a drop of 0.263. The combined post-correction worsens the average performance and increases the sensitivity of the descriptive metrics to single-node removal in this case.

Appendix D. Supplementary statistical and diagnostic results

D.1. Spatial-dependence diagnostics and permutation details

This subsection collects the diagnostics and transparency tables for the inference protocol of Section 4.3, computed for all 27 registered comparison–metric pairs of the British main case (9 comparisons $\times$ 3 metrics). The main spatial weight matrix is queen contiguity over the 16

study regions (neighbor counts range from 1 to 6; no region is isolated), with k -nearest-neighbor ($k = 3$) weights as a supplement; Moran's I significance is assessed using 999 random permutations, two-sided. Where the queen diagnostic is significant at the 0.05, the protocol replaces the primary p -value by a spatial block permutation: the 16 regions are greedily matched into 8 contiguous blocks, paired differences are summed within blocks, and signs are flipped at the block level with exact enumeration of all $2^8 = 256$ assignments, so the smallest attainable two-sided p is $2/256 \approx 0.008$. Table D.21 reports all diagnostics; 8 of the 27 pairs trigger the replacement. Table D.22 reports the per-seed exact sign-flip permutation p -values for every comparison involving a GNN-based configuration, together with the Cauchy combination as a reference column; the primary value in the main text remains the permutation p computed after seed averaging, whose Holm-corrected value (within each nine-comparison metric family) the last column reports, and static–static comparisons are deterministic and carry no seed dimension.

Australian queen-contiguity diagnostics cover the same nine registered comparisons and three metrics. Three correlation contrasts are flagged: UniNP versus Uni ($I = 0.426$, permutation $p = 0.042$), GPMpostNP versus GPM ($I = 0.428$, $p = 0.049$), and GPMpostNP versus GPMpostP ($I = 0.189$, $p = 0.034$). Their contiguous-block permutation values are 0.094, 0.125, and 0.148, respectively. No Australian RMSE comparison is flagged, and no reported significance verdicts change.

D.2. Post-registry additive comparison: GNNaddP versus GPMpostNP

Table D.23 reports the tests behind the primary comparison in Section 5.1 between GNNaddP and GPMpostNP. The additive configurations were trained after the British comparison registry was frozen, so these tests form their own post-registry family: seed-averaged paired region differences, exact two-sided sign-flip permutation (16 regions, 2^{16} enumeration), Holm-corrected within one family of 14 post-registry comparisons per metric.

Table D.21

Moran's I spatial-dependence diagnostics for all 27 registered comparison–metric pairs (British main case). I and two-sided permutation p (999 permutations) under queen contiguity and k -nearest-neighbor ($k = 3$) weights; the last column gives the spatial block-permutation p (8 contiguous blocks, exact 2^8 enumeration), computed when the queen diagnostic is significant at 0.05.

ID	Comparison	Metric	I_{queen}	p_{queen}	I_{kNN3}	p_{kNN3}	Block-perm. p
1	GNN vs GPM	RMSE	0.113	0.407	0.248	0.062	—
1	GNN vs GPM	MAE	0.203	0.180	0.212	0.094	—
1	GNN vs GPM	Corr	0.252	0.100	0.145	0.227	—
2	GNNpostP vs GNN	RMSE	0.294	0.062	0.294	0.030	—
2	GNNpostP vs GNN	MAE	0.136	0.335	0.191	0.129	—
2	GNNpostP vs GNN	Corr	0.177	0.234	0.145	0.238	—
3	GNNpostN vs GNN	RMSE	0.637	0.002	0.533	0.004	0.562
3	GNNpostN vs GNN	MAE	0.473	0.011	0.445	0.006	0.570
3	GNNpostN vs GNN	Corr	0.035	0.623	0.042	0.532	—
4	GNNpostNP vs GNNpostP	RMSE	0.341	0.033	0.320	0.030	0.008
4	GNNpostNP vs GNNpostP	MAE	0.225	0.112	0.210	0.093	—
4	GNNpostNP vs GNNpostP	Corr	0.222	0.146	0.151	0.187	—
5	GPMpostNP vs GPMpostN	RMSE	0.329	0.042	0.314	0.012	0.008
5	GPMpostNP vs GPMpostN	MAE	0.263	0.107	0.284	0.034	—
5	GPMpostNP vs GPMpostN	Corr	0.092	0.427	0.044	0.511	—
6	GPMpostP vs GPM	RMSE	0.302	0.029	0.265	0.022	0.008
6	GPMpostP vs GPM	MAE	0.255	0.096	0.274	0.032	—
6	GPMpostP vs GPM	Corr	0.160	0.281	0.183	0.135	—
7	GPMpostNP vs GPMpostP	RMSE	0.325	0.063	0.169	0.139	—
7	GPMpostNP vs GPMpostP	MAE	0.301	0.051	0.253	0.049	—
7	GPMpostNP vs GPMpostP	Corr	0.303	0.046	0.144	0.217	0.008
8	GNNpriorN vs GNN	RMSE	0.277	0.080	0.333	0.012	—
8	GNNpriorN vs GNN	MAE	0.225	0.136	0.311	0.029	—
8	GNNpriorN vs GNN	Corr	0.281	0.069	0.288	0.021	—
9	GNNpostP vs GPMpostNP	RMSE	0.447	0.002	0.249	0.051	0.008
9	GNNpostP vs GPMpostNP	MAE	0.651	0.002	0.292	0.028	0.008
9	GNNpostP vs GPMpostNP	Corr	0.229	0.129	0.043	0.511	—

Table D.22

Per-seed exact sign-flip permutation p -values for the comparisons involving GNN-based configurations (6 comparisons $\times$ 3 metrics), with the Cauchy combination as a reference column. The primary value in the main text is the permutation p after seed averaging; the last column gives its Holm-corrected value within each nine-comparison metric family. Static-static comparisons (IDs 5–7) are deterministic and have no seed dimension.

ID	Comparison	Metric	$p_{\text{seed } 42}$	$p_{\text{seed } 123}$	$p_{\text{seed } 456}$	Cauchy	Holm p
1	GNN vs GPM	RMSE	1.2×10^{-4}	1.8×10^{-4}	1.2×10^{-4}	1.4×10^{-4}	4.9×10^{-4}
1	GNN vs GPM	MAE	1.8×10^{-4}	3.1×10^{-4}	1.2×10^{-4}	1.8×10^{-4}	9.2×10^{-4}
1	GNN vs GPM	Corr	2.4×10^{-4}	9.2×10^{-5}	1.5×10^{-4}	1.4×10^{-4}	6.1×10^{-4}
2	GNNpostP vs GNN	RMSE	0.540	0.826	0.652	0.708	1.000
2	GNNpostP vs GNN	MAE	0.584	0.169	0.594	0.386	1.000
2	GNNpostP vs GNN	Corr	0.170	0.062	0.073	0.084	0.273
3	GNNpostN vs GNN	RMSE	0.422	0.497	0.478	0.465	1.000
3	GNNpostN vs GNN	MAE	0.777	0.149	0.254	0.330	1.000
3	GNNpostN vs GNN	Corr	1.2×10^{-4}	6.1×10^{-5}	3.1×10^{-5}	5.2×10^{-5}	5.5×10^{-4}
4	GNNpostNP vs GNNpostP	RMSE	3.1×10^{-5}	3.1×10^{-5}	6.1×10^{-5}	3.7×10^{-5}	2.7×10^{-4}
4	GNNpostNP vs GNNpostP	MAE	6.1×10^{-5}	6.1×10^{-5}	3.1×10^{-5}	4.6×10^{-5}	4.3×10^{-4}
4	GNNpostNP vs GNNpostP	Corr	0.770	0.783	0.882	0.826	1.000
8	GNNpriorN vs GNN	RMSE	0.754	0.255	0.862	0.702	1.000
8	GNNpriorN vs GNN	MAE	0.326	0.343	0.715	0.464	1.000
8	GNNpriorN vs GNN	Corr	0.392	0.018	0.004	0.011	0.123
9	GNNpostP vs GPMpostNP	RMSE	3.1×10^{-5}	3.1×10^{-5}	3.1×10^{-5}	3.1×10^{-5}	2.7×10^{-4}
9	GNNpostP vs GPMpostNP	MAE	1.5×10^{-4}	2.4×10^{-4}	3.1×10^{-4}	2.2×10^{-4}	9.8×10^{-4}
9	GNNpostP vs GPMpostNP	Corr	0.363	0.834	0.538	0.639	1.000

Table D.23

GNNaddP versus GPMpostNP: seed-averaged mean paired difference and Holm-corrected exact sign-flip permutation p , Holm-corrected within the post-registry additive family of 14 comparisons per metric (family identifiers from the frozen artifact).

Metric	Mean diff.	Holm p	Holm family
RMSE	-0.28	0.163	r12_main_rmse
MAE	-0.24	0.216	r12_main_mae
Corr	+0.073	0.011	r12_main_corr

D.3. Australian supplementary-configuration tests

Table D.24 collects the tests behind the Australian feature-fusion and single-prior statements of Sections 5.1, Section 5.4, and Section 5.10. These configurations were trained after the Australian registry freeze, so every p -value is the raw exact two-sided sign-flip permutation p (12 paired SA4 regions, seeds averaged first, 2^{12} enumeration) and enters no Holm family; the GNNpriorN RMSE row also appears as comparison #8 in the nine-comparison Australian family of Table 4, where its Holm-corrected value coincides with the raw 0.956 reported here.

Table D.24

Australian post-registry configuration tests against the GNN base: mean paired difference and raw exact sign-flip permutation p (12 regions; raw values, outside any Holm family, because the runs postdate the registry).

Configuration	Metric	Mean diff.	Raw p
GNNfusion	RMSE	-7.67	0.015
GNNfusion	MAE	-4.05	0.024
GNNfusion	Corr	+0.095	0.004
GNNpriorN	RMSE	+1.01	0.956
GNNpriorP	RMSE	-0.04	0.983
GNNpriorNP	RMSE	-0.003	0.994

D.4. Probing suite: full results and CKA

Table D.25 reports the complete probing suite from Section 5.4 under both evaluation splits. Each probe is a standardized two-layer perceptron (one hidden layer of 64 ReLU units) regressing the log-transformed, per-region standardized signal from the 128-dimensional embeddings. The region-blocked split is the primary split: probe training and testing never share a region. Random splits, which mix grid points of the same region across training and testing, inflate the base-embedding score from 0.074 to 0.452 for NTL (and from 0.040 to 0.175 for Proximity), a factor of roughly six that measures spatial leakage rather than encoded signal, which is why single-split probing can overstate what a representation contains. The paired fusion-versus-base gains on the region-blocked split are +0.94 for NTL (bootstrap CI [0.69, 1.36]) and +0.95 for Proximity (CI [0.82, 1.14]), each with sign-flip $p = 3.1 \times 10^{-5}$. Table D.26 corroborates the probe with a linear similarity measure: the CKA between base embeddings and either signal is of order 10^{-3} for every seed, compared to 0.11 for the raw land-use features and 0.80 for the positive-control inputs.

Table D.25

Full probing suite: out-of-region (region-blocked) and random-split pooled R^2 for both targets. Each metric cell reports region-blocked / random-split values. The random split is reported only as an inflation contrast; the region-blocked value is the primary value (cf. Table 8). Rows (i) and (ii) are the target-containing positive control and the signal-injected fusion representation, respectively.

Representation	NTL	Prox
(i) Positive control	0.99997/0.99998	0.99998/0.99998
(ii) Fusion embeddings	0.988/0.998	0.973/0.990
(iii) Baseline embeddings	0.074/0.452	0.040/0.175
(iv) Raw land-use features	0.357/0.383	0.136/0.143

Table D.26

Linear representation similarity (CKA) between each representation and the auxiliary signals.

Representation	CKA NTL	CKA Prox
Positive-control inputs (7-dim)	0.805	0.797
Raw land-use features (5-dim)	0.108	0.042
Baseline embeddings (seed 42)	0.0009	0.0004
Baseline embeddings (seed 123)	0.0006	0.0003
Baseline embeddings (seed 456)	0.0010	0.0004

D.5. Scale-free accuracy metrics

Table D.28 supplements the scale-dependent errors in **Table 3**. WAPE is $\sum_j |y_j - \hat{y}_j| / \sum_j |y_j|$, and mean-normalized RMSE is $\text{RMSE} / \text{mean}_j |y_j|$. Both are computed within each region under the same seed-handling protocol as in the main table. EqualStation uses D_r/n_r ; direct EqualSubregion uses D_s/n_s at the substation output. MAPE is not reported: five British regions contain six substations with zero measured demand, causing 220 of the 1273 evaluation rows to be undefined, and a single substation contributes up to 66.5% of the non-zero APE sum in the most concentrated row.

Table D.27 separates the two pipeline-external accuracy references from the Voronoi-pipeline evaluation. EqualStation is the predictive- R^2 zero point and has undefined within-region correlation because its predictions are constant. Direct EqualSubregion has pooled predictive $R^2 = 0.248$ in Britain and 0.194 in Australia, with mean correlations of 0.327 and 0.477. It answers an output-layer information-parity question, whereas the main table asks whether weighting and correction add value at fixed grid support.

D.6. EqualSubregionGrid reference configurations

Table D.29 reports all seven EqualSubregionGrid configurations for Britain, Australia, and Germany. The construction was added after the original evaluation was complete and is reported as a post-hoc robustness analysis. It shares the grid support, D_s conservation, frozen Voronoi assignment, and correction operators used by the evaluated configurations. It is not treated as a preregistered method group.

D.7. Regime-test confusion matrices

Table D.30 decomposes the configuration-level RMSE-sign classification in **Section 5.5**. The rows are not independent observations: Britain contributes 240 configurations nested in 16 regions, Australia 180 in 12

Table D.28

Scale-free error metrics for the two main study cases (region mean $\pm$ inter-region standard deviation). Lower values are better. EqualStation and direct EqualSubregion are pipeline-external references based on D_r/n_r and D_s/n_s , respectively.

Configuration	WAPE		Mean-normalized RMSE	
	GB	AU	GB	AU
EqualStation	40.8 $\pm$ 13.3%	34.6 $\pm$ 14.5%	0.540 $\pm$ 0.174	0.419 $\pm$ 0.180
EqualSubreg.	37.5 $\pm$ 10.6%	27.5 $\pm$ 12.9%	0.499 $\pm$ 0.145	0.350 $\pm$ 0.164
Uni	68.5 $\pm$ 9.8%	53.9 $\pm$ 25.0%	0.967 $\pm$ 0.173	0.733 $\pm$ 0.369
UniN	58.9 $\pm$ 8.2%	50.3 $\pm$ 21.3%	0.821 $\pm$ 0.135	0.677 $\pm$ 0.294
UniP	49.8 $\pm$ 8.8%	36.2 $\pm$ 14.7%	0.662 $\pm$ 0.121	0.456 $\pm$ 0.180
UniNP	40.7 $\pm$ 7.1%	34.0 $\pm$ 13.7%	0.546 $\pm$ 0.105	0.425 $\pm$ 0.167
GPM	65.2 $\pm$ 11.2%	54.9 $\pm$ 24.8%	0.921 $\pm$ 0.187	0.757 $\pm$ 0.366
GPMpostN	56.3 $\pm$ 9.0%	51.6 $\pm$ 22.1%	0.781 $\pm$ 0.135	0.699 $\pm$ 0.300
GPMpostP	48.1 $\pm$ 9.0%	37.4 $\pm$ 18.1%	0.634 $\pm$ 0.117	0.478 $\pm$ 0.219
GPMpostNP	40.4 $\pm$ 7.0%	35.5 $\pm$ 18.0%	0.539 $\pm$ 0.097	0.451 $\pm$ 0.216
GPMaddN	55.9 $\pm$ 10.3%	51.3 $\pm$ 23.5%	0.769 $\pm$ 0.146	0.694 $\pm$ 0.335
GPMaddP	57.3 $\pm$ 11.8%	50.8 $\pm$ 23.2%	0.790 $\pm$ 0.161	0.697 $\pm$ 0.344
GPMaddNP	56.6 $\pm$ 11.9%	50.7 $\pm$ 23.2%	0.785 $\pm$ 0.165	0.695 $\pm$ 0.343
GNN	49.1 $\pm$ 4.1%	66.0 $\pm$ 26.1%	0.680 $\pm$ 0.056	0.931 $\pm$ 0.589
GNNpostN	48.6 $\pm$ 3.3%	63.5 $\pm$ 25.2%	0.696 $\pm$ 0.078	0.892 $\pm$ 0.582
GNNpostP	48.9 $\pm$ 4.8%	45.3 $\pm$ 25.9%	0.699 $\pm$ 0.090	0.648 $\pm$ 0.550
GNNpostNP	57.1 $\pm$ 8.0%	44.3 $\pm$ 26.1%	0.838 $\pm$ 0.158	0.637 $\pm$ 0.554
GNNaddN	46.6 $\pm$ 4.3%	54.8 $\pm$ 28.8%	0.647 $\pm$ 0.071	0.741 $\pm$ 0.627
GNNaddP	38.5 $\pm$ 6.0%	37.7 $\pm$ 17.4%	0.517 $\pm$ 0.087	0.487 $\pm$ 0.239
GNNaddNP	39.8 $\pm$ 4.5%	37.0 $\pm$ 17.1%	0.542 $\pm$ 0.076	0.476 $\pm$ 0.238
GNNpriorN	48.7 $\pm$ 3.8%	66.1 $\pm$ 30.4%	0.668 $\pm$ 0.063	0.944 $\pm$ 0.692
GNNpriorP	48.4 $\pm$ 3.6%	64.8 $\pm$ 28.6%	0.666 $\pm$ 0.060	0.924 $\pm$ 0.637
GNNpriorNP	47.7 $\pm$ 3.3%	65.6 $\pm$ 27.7%	0.658 $\pm$ 0.055	0.926 $\pm$ 0.620
GNNfusion	60.6 $\pm$ 8.3%	57.1 $\pm$ 22.3%	0.835 $\pm$ 0.134	0.775 $\pm$ 0.445

regions, and Germany 15 in one region. Region-cluster bootstrap intervals are 75.4–90.8% for Britain and 96.7–100% for Australia; no interval is interpreted for Germany. The false-positive concentration in Britain (34 FP versus 6 FN) is reported as an empirical asymmetry, not converted into a cross-case-transfer claim.

D.8. Isoline percentile distributions

Fig. D.9 shows a finite Monte Carlo reference distribution for each of the 48 (region, seed) units. Each Dirichlet draw lies exactly on the reconstruction-loss isoline because per-class weight masses are fixed within every sub-region, and a hit-and-run sampler serves as a mixing control. The 200 draws do not exhaust the isoline, establish the feasible-set extrema, or guarantee reachability under the shared GNN parameterization.

Table D.27

Pipeline-external absolute-accuracy references. EqualStation uses only the regional total D_r and is the exact predictive- R^2 zero point. Direct EqualSubregion assigns D_s/n_s at the substation output and does not traverse the grid/Voronoi pipeline. Cells report region mean $\pm$ inter-region standard deviation; predictive- R^2 cells give the within-region pooled value in parentheses. These rows provide accuracy context and are not competitors in the controlled two-axis evaluation.

Reference	RMSE		MAE		Corr		Predictive R^2	
	GB	AU	GB	AU	GB	AU	GB	AU
EqualStation	7.46 $\pm$ 3.64	16.80 $\pm$ 8.53	5.62 $\pm$ 2.74	13.84 $\pm$ 6.64	undefined	undefined	0.000 $\pm$ 0.000 (0.000)	0.000 $\pm$ 0.000 (0.000)
Direct EqualSubregion	6.80 $\pm$ 2.84	13.93 $\pm$ 9.03	5.10 $\pm$ 1.99	10.92 $\pm$ 6.86	0.327 $\pm$ 0.145	0.477 $\pm$ 0.283	0.128 $\pm$ 0.095 (+0.248)	0.307 $\pm$ 0.274 (+0.194)

Table D.29

Complete EqualSubregionGrid reference family. Britain and Australia report the regional mean $\pm$ inter-region standard deviation; Germany is a single descriptive region. Predictive- R^2 cells give the within-region pooled value in parentheses. The seven configurations are the base and the full multiplicative (post) and additive correction sets; no configuration is selected after inspecting its results. WAPE is reported as a percentage.

Study case	Grid config.	RMSE	MAE	Corr	Predictive R^2	WAPE
GB	Base	8.59 $\pm$ 2.98	6.50 $\pm$ 2.00	0.143 $\pm$ 0.141	-0.493 $\pm$ 0.379 (-0.142)	48.0 $\pm$ 8.9%
	postN	8.22 $\pm$ 2.82	6.16 $\pm$ 1.95	0.245 $\pm$ 0.168	-0.403 $\pm$ 0.445 (-0.049)	45.3 $\pm$ 7.0%
	postP	9.40 $\pm$ 3.08	6.97 $\pm$ 2.19	0.169 $\pm$ 0.193	-0.905 $\pm$ 0.893 (-0.259)	51.3 $\pm$ 9.3%
	postNP	10.49 $\pm$ 3.18	7.56 $\pm$ 2.23	0.210 $\pm$ 0.202	-1.503 $\pm$ 1.460 (-0.499)	55.6 $\pm$ 8.2%
	addN	8.34 $\pm$ 2.80	6.30 $\pm$ 2.07	0.419 $\pm$ 0.189	-0.521 $\pm$ 0.683 (-0.096)	46.0 $\pm$ 5.0%
	addP	8.20 $\pm$ 2.85	6.21 $\pm$ 2.00	0.221 $\pm$ 0.156	-0.364 $\pm$ 0.358 (-0.018)	45.9 $\pm$ 10.2%
	addNP	8.14 $\pm$ 2.72	6.20 $\pm$ 1.88	0.319 $\pm$ 0.185	-0.407 $\pm$ 0.524 (-0.009)	45.7 $\pm$ 7.3%
AU	Base	21.74 $\pm$ 8.82	17.78 $\pm$ 7.10	0.048 $\pm$ 0.472	-1.137 $\pm$ 1.789 (-0.368)	42.5 $\pm$ 9.9%
	postN	21.39 $\pm$ 8.98	17.55 $\pm$ 7.21	0.091 $\pm$ 0.468	-1.063 $\pm$ 1.797 (-0.329)	41.7 $\pm$ 9.7%
	postP	21.36 $\pm$ 9.59	16.93 $\pm$ 7.90	0.260 $\pm$ 0.399	-1.214 $\pm$ 2.065 (-0.361)	39.9 $\pm$ 12.2%
	postNP	21.54 $\pm$ 9.62	17.11 $\pm$ 7.88	0.276 $\pm$ 0.405	-1.278 $\pm$ 2.149 (-0.379)	40.4 $\pm$ 12.3%
	addN	21.06 $\pm$ 8.84	16.99 $\pm$ 7.07	0.138 $\pm$ 0.533	-0.950 $\pm$ 1.528 (-0.313)	40.4 $\pm$ 9.6%
	addP	20.10 $\pm$ 9.00	16.07 $\pm$ 7.37	0.200 $\pm$ 0.445	-0.954 $\pm$ 1.956 (-0.164)	37.4 $\pm$ 9.3%
	addNP	20.03 $\pm$ 8.98	15.96 $\pm$ 7.37	0.223 $\pm$ 0.446	-0.954 $\pm$ 1.960 (-0.157)	37.2 $\pm$ 9.1%
DE	Base	5.79	4.81	0.876	0.710 (+0.710)	30.1%
	postN	6.23	5.41	0.859	0.665 (+0.665)	33.9%
	postP	6.26	5.11	0.860	0.662 (+0.662)	32.0%
	postNP	7.39	5.22	0.821	0.528 (+0.528)	32.7%
	addN	6.29	5.31	0.858	0.659 (+0.659)	33.3%
	addP	6.00	5.21	0.863	0.689 (+0.689)	32.6%
	addNP	6.31	5.39	0.853	0.656 (+0.656)	33.7%

Table D.30

Configuration-level confusion matrices for original-scale RMSE direction (positive class = correction helps). The 435 rows are repeated configurations nested in 29 region clusters; accuracies therefore describe evaluated configurations rather than independent study cases. Base rate denotes the accuracy of always predicting the majority class and differs from the always-apply rate when correction helps is not the majority class. The lower panel stratifies rows by their absolute margin from the strict boundary; the 0.195 band is the maximum simulated displacement after Voronoi aggregation and change to the original-scale RMSE.

Case/stratum	n	TP	FP	FN	TN	Acc.	Maj. rate
Britain	240	145	34	6	55	0.833	0.629
Australia	180	167	2	1	10	0.983	0.933
Germany	15	6	1	0	8	0.933	0.600
All	435	—	—	—	—	0.899	0.747
<i>Boundary-margin strata</i>							
Outside (> 0.195)	331	—	—	—	—	0.970	0.831
Inside (≤ 0.195)	104	—	—	—	—	0.673	0.519
All rows	435	—	—	—	—	0.899	0.747

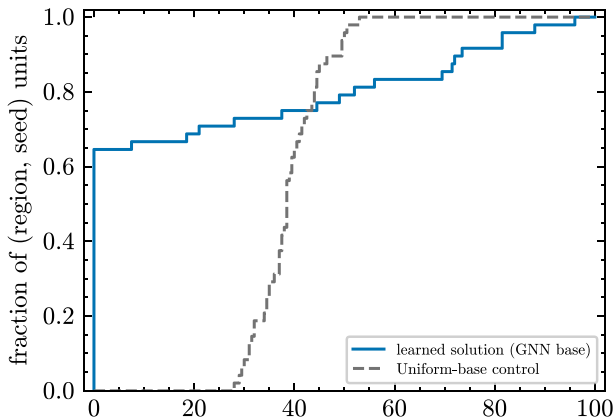


Fig. D.9. Empirical distribution functions of percentiles within the specified 200-draw Monte Carlo isoline reference for 48 (region, seed) units. The learned solution has a median sampled percentile of 0.00% and outperforms all draws in 31 units; the Uniform-base control has a median of 39%. These are finite-sample reference percentiles, not bounds on the full isoline.

Appendix E. Propositions and proofs

E.1. Proof of Proposition 1

Within a fixed sub-region s , the denominator of Eq. (11), $\sum_{a' \in \mathcal{A}_s} d_{a'}^{\text{base}} f(a')$, does not depend on a . Eq. (11) is therefore the map $d_a^{\text{base}} f(a) \mapsto C_s \cdot d_a^{\text{base}} f(a)$ with the single positive constant $C_s = D_s / \sum_{a'} d_{a'}^{\text{base}} f(a')$, which is Eq. (12). Claim (i) follows because a common positive factor cancels from every ratio $d_a^{\text{corr}} / d_{a'}^{\text{corr}}$. Claim (ii) follows because the no-re-normalization variant returns $d_a^{\text{base}} f(a)$ itself, which differs from Eq. (11) exactly by C_s on each sub-region.

E.2. Proof of Proposition 2

Assume $\sigma_{\text{res}} > 0$ and $\sigma_c > 0$; if $\sigma_c = 0$ the factor is constant, the correction is neutral, and there is nothing to prove.

We first consider geometric re-normalization. Multiplying each base estimate by its factor means dividing all products $b_i f_i$ by the constant $G = \exp(\frac{1}{n} \sum_i \log f_i)$, so that the mean log estimate is preserved. The corrected estimate b'_i therefore satisfies

$$\log b'_i = \log b_i + \log f_i - \frac{1}{n} \sum_j \log f_j = \log b_i + c_i, \quad (\text{E.1})$$

with c_i the centered log factor of Proposition 2, satisfying $\sum_i c_i = 0$.

We next compare the errors. The log residual of the corrected estimate is $r'_i = \log(y_i / b'_i) = r_i - c_i$. Hence

$$\sum_i (r_i - c_i)^2 < \sum_i r_i^2 \iff 2 \langle \mathbf{r}, \mathbf{c} \rangle > \|\mathbf{c}\|^2, \quad (\text{E.2})$$

by expanding the square; the inequality is strict on both sides, which gives the first form of the criterion.

Finally, the correlation form follows because $\sum_i c_i = 0$, so the inner product loses no generality by centering r :

$$\langle \mathbf{r}, \mathbf{c} \rangle = \sum_i (r_i - \bar{r}) c_i + \bar{r} \sum_i c_i = \sum_i (r_i - \bar{r})(c_i - \bar{c}) = n \rho \sigma_{\text{res}} \sigma_c, \quad (\text{E.3})$$

where $\bar{r}, \bar{c}$ are the means and the last step uses the population definitions of correlation and standard deviation. With $\|\mathbf{c}\|^2 = n \sigma_c^2$, the first form of Eq. (13) becomes $2n \rho \sigma_{\text{res}} \sigma_c > n \sigma_c^2$, i.e., $\rho > \sigma_c / (2 \sigma_{\text{res}})$. No assumption on the mean of r is required: the centering of c absorbs it.

Table E.31

Sensitivity of the 435-configuration criterion counts to the log floor and zero-demand substations. The default protocol is all substations with $\epsilon = 10^{-6} \sum_j y_j$. “Side flips” counts predicted-side changes relative to that default among the 240 British rows; “RMSE-label flips” counts changes in the British RMSE direction after zero-demand substations are excluded. The final column reports the mean residual dispersion σ_{res} .

Substation set	$\epsilon / \sum_j y_j$	Correct log-MSE	Correct RMSE	Side flips	RMSE-label flips	Mean σ_{res}
All substations	10^{-8}	425/435 (97.7%)	389/435 (89.4%)	2	0	1.103
All substations	10^{-6}	425/435 (97.7%)	391/435 (89.9%)	0	0	0.969
All substations	10^{-4}	429/435 (98.6%)	396/435 (91.0%)	5	0	0.841
Positive truth only	10^{-8}	427/435 (98.2%)	395/435 (90.8%)	9	3	0.812
Positive truth only	10^{-6}	427/435 (98.2%)	395/435 (90.8%)	9	3	0.812
Positive truth only	10^{-4}	430/435 (98.9%)	396/435 (91.0%)	10	3	0.793

The exact proof above ends with geometric re-normalization and log-MSE. The implemented correction operator instead conserves mass arithmetically, and the principal empirical label is the sign of the RMSE change on the original demand scale after Voronoi aggregation. The following simulation therefore records the sensitivity of the boundary; it is not an extension of the proposition.

Synthetic allocations of 2000 grid points were evaluated over 24 alignment values ($\rho = -0.2, \dots, 0.95$), 16 dispersion ratios ($\sigma_c/\sigma_{\text{res}} = 0.1, \dots, 1.6$), and 60 replicates per design point. Under the strict geometric/log-MSE correction operator, the maximum absolute deviation from $\rho = \sigma_c/(2\sigma_{\text{res}})$ is 0.00587. Arithmetic mass conservation increases the maximum deviation to 0.07186 while retaining the log-MSE. When the loss is changed to the demand-space MSE, the maximum identifiable deviation is 0.19337.

After Voronoi aggregation (Layer B), an RMSE boundary is identifiable for 15 of the 16 dispersion ratios. Its maximum absolute deviation is 0.18929 when estimated from the fraction of improving replicates and 0.19487 when estimated from the mean RMSE difference. These deviations define the displacement band, the interval of margins $|\rho - \sigma_c/(2\sigma_{\text{res}})| \leq 0.195$ within which the strict boundary is not claimed to hold. They motivate reporting the threshold as a fixed-rule classifier under the stated floor protocol, while preserving the if-and-only-if language only for the first, geometric/log-MSE row.

Data availability

All data are open source at <https://github.com/KIT-IAI/AllocateGNN>

References

- Hörsch J, Hofmann F, Schlachtberger D, Brown T. PyPSA-Eur: an open optimisation model of the European transmission system. *Energy Strategy Rev* 2018;22:207–15. <https://doi.org/10.1016/j.esr.2018.08.012>
- Hüll L, Wienholt L, Cußmann I, Müller UP, Matke C, Köttler E. Allocation of annual electricity consumption and power generation capacities across multiple voltage levels in a high spatial resolution. *Int J Sustain Energy Plan Manag* 2017;13:79–92. <https://doi.org/10.5278/ijsep.2017.13.6>
- Raventós O, Dengiz T, Medjroubi W, Unaichi C, Bruckmeier A, Finck R. Comparison of different methods of spatial disaggregation of electricity generation and consumption time series. *Renew Sustain Energy Rev* 2022;163:112186. <https://doi.org/10.1016/j.rser.2022.112186>
- Monteiro J, Martins B, Murrieta-Flores P, Pires J. Spatial disaggregation of historical census data leveraging multiple sources of ancillary information. *ISPRS Int. J. Geo Inf* 2019;8:327. <https://doi.org/10.3390/ijgi8080327>
- Sapena M, Kühnl M, Wurm M, Patiño JE, Duque J, Taubenböck H. Empiric recommendations for population disaggregation under different data scenarios. *PLoS One* 2022;17. <https://doi.org/10.1371/journal.pone.0274504>
- Qiu Y, Zhao X, Fan D, Li S, Zhao Y. Disaggregating population data for assessing progress of SDGs: methods and applications. *Int J Digit Earth* 2022;15:2–29. <https://doi.org/10.1080/17538947.2021.2013553>
- Crato N. Figuring it out. Berlin, Heidelberg: Springer; 2010. <https://doi.org/10.1007/978-3-642-04833-3>
- Chen J, Gao M, Hou W, Song M, Liu X, Yu L. Global 1 KM ×1 KM gridded revised real gross domestic product and electricity consumption during 1992–2019 based on calibrated nighttime light data. *Sci Data* 2022;9:202. <https://doi.org/10.1038/s41597-022-01322-5>
- Borges CE, Kamara-Esteban O, Castillo-Calzadilla T, Andonegui CM, Alonso-Vicario A. Enhancing the missing data imputation of primary substation load demand records. *Sustain Energy Grids Netw* 2020;23:100369. <https://doi.org/10.1016/j.segan.2020.100369>
- Patil S, Pflugradt N, Weinand JM, Stolten D, Kropp J. A systematic review of spatial disaggregation methods for climate action planning. *Energy AI* 2024;17:100386. <https://doi.org/10.1016/j.egyai.2024.100386>
- Mu X, Geiges J, Liu N, Schlachter T, Hagenmeyer V. Improving spatial allocation for energy system coupling with graph neural networks. *Electr Power Syst Res* 2026;262:113519. <https://doi.org/10.1016/j.epsr.2026.113519>. <https://www.sciencedirect.com/science/article/pii/S0378779626008126>
- Singh A, Eser P, Chokani N, Abhari R. High resolution modeling of the impacts of exogenous factors on power systems—case study of Germany. *Energies* 2015;8. <https://doi.org/10.3390/en81212424>
- Chen H, Flores GEC, Li C. Physics-informed neural networks with hard linear equality constraints. *Comput Chem Eng* 2024;189:108764. <https://doi.org/10.1016/j.compchemeng.2024.108764>
- Liang K, Liu Y, Zhou S, Tu W, Wen Y, Yang X, Dong X, Liu X. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Trans Knowl Data Eng* 2024;36:226–38. <https://doi.org/10.1109/TKDE.2023.3282989>
- Gong T, Lee T, Stephenson C, Renduchintala V, Padhy S, Ndirango A, Keskin G, Elibol OH. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access* 2019;7:141627–32. <https://doi.org/10.1109/ACCESS.2019.2943604>
- Cochran WG ed. Sampling techniques. 3 ed. New York: John Wiley & Sons; 1977.
- Lavenberg SS, Welch PD. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Manag Sci* 1981;27:322–35. <https://doi.org/10.1287/mnsc.27.3.322>
- Aryanpur V, Ógallachóir B, Dai H, Chen W, Glynn J. A review of spatial resolution and regionalisation in national-scale energy systems optimisation models. *Energy Strategy Rev* 2021;37:100702. <https://doi.org/10.1016/j.esr.2021.100702>
- Frysztacki MM, Hörsch J, Hagenmeyer V, Brown T. The strong effect of network resolution on electricity system models with high shares of wind and solar. *Appl Energy* 2021;291:116726. <https://doi.org/10.1016/j.apenergy.2021.116726>
- Jalil-Vega F, Hawkes A. The effect of spatial resolution on outcomes from energy systems modelling of heat decarbonisation. *Energy* 2018;155:339–50. <https://doi.org/10.1016/j.energy.2018.04.160>
- Martínez-Gordón R, Morales-España G, Sijm J, Faaij A. A review of the role of spatial resolution in energy systems modelling: lessons learned and applicability to the North sea region. *Renew. & Sustain. Energy Rev* 2021;141:110857. <https://doi.org/10.1016/j.rser.2021.110857>
- Patil S, Pflugradt N, Weinand J, Kropp J, Stolten D. Spatially disaggregated energy consumption and emissions in end-use sectors for Germany and Spain. *arXiv:2505.05139* [arXiv preprint]. 2025.
- Mei J, Goude Y, Hebrail G, Kong N. Spatial estimation of electricity consumption using socio-demographic information. In: 2016 IEEE PES Asia-Pacific power and energy engineering conference (APPEEC); 2016. p. 753–7. <https://doi.org/10.1109/APPEEC.2016.7779596>
- Köppel S, Böing F, Pellingner C. Modeling of the transmission grid using geo allocation and generalized processes. In: Bertsch V, Fichtner W, Heuveline V, Leibfried T, editors. *Advances in energy system optimization*. Springer International Publishing; 2017. p. 189–204. https://doi.org/10.1007/978-3-319-51795-7_12
- Georgati M, Monteiro J, Martins B, Keffler C, Hansen HS. Modeling population distribution: a visual and quantitative analysis of global boosting and deep learning models for multi-output spatial disaggregation. *Trans. GIS* 2024;28:130–53. <https://doi.org/10.1111/tgis.13130>
- Monteiro J, Martins B, Pires J. A hybrid approach for the spatial disaggregation of socio-economic indicators. *Int J Data Sci Anal* 2018;5:189–211. <https://doi.org/10.1007/s41060-017-0080-z>
- Bhattarai D, Luciere A, Lovell H, Aryal J. Remote sensing of night-time lights and electricity consumption: a systematic literature review and meta-analysis. *Geogr Compass* 2023;17:e12684. <https://doi.org/10.1111/gec3.12684>
- Crippa M, Guizzardi D, Pagani F, Schiavina M, Melchiorri M, Pisoni E, Graziosi F, Muntean M, Maes J, Dijkstra L, Van Damme M, Clarisse L, Coheur P. Insights into the spatial distribution of global, national, and subnational greenhouse gas emissions in the emissions database for global atmospheric research (EDGAR v8.0). *Earth Syst Sci Data* 2024;16:2811–30. <https://doi.org/10.5194/essd-16-2811-2024>
- Danylo O, Bun R, See L, Charkovska N. High-resolution spatial distribution of greenhouse gas emissions in the residential sector. *Mitig Adapt Strateg Glob Change* 2019;24:941–67. <https://doi.org/10.1007/s11027-019-9846-z>

- [30] Baran M, Freeman L, Hanson F, Ayers V. Load estimation for load monitoring at distribution substations. *IEEE Trans Power Syst* 2005;20:164–70. <https://doi.org/10.1109/TPWRS.2004.840409>
- [31] Mu X, Geiges J, Liu J, Schlachter T, Hagenmeyer V. Improving spatial allocation for energy system coupling with clustering-based voronoi diagrams. In: 2025 IEEE 13th international conference on smart energy grid engineering (SEGE); 2025. p. 195–200. <https://doi.org/10.1109/SEGE65970.2025.11203477>
- [32] Hu Z, Dong Y, Wang K, Sun Y. Heterogeneous graph transformer. In: Proceedings of the web conference 2020 (WWW '20). New York, NY, USA: Association for Computing Machinery; 2020. p. 2704–10. <https://doi.org/10.1145/3366423.3380027>
- [33] Najafi Ravadanegh S, Gholizadeh Roshanagh R. A heuristic algorithm for optimal multistage sizing, siting and timing of MV distribution substations. *Electr Power Syst Res* 2013;105:134–41. <https://doi.org/10.1016/j.epsr.2013.07.019>
- [34] Bosio A, Berizzi A, Merlo M, Morotti A, Iannarelli G. A GIS-based approach for primary substations siting and timing based on voronoi diagram and particle swarm optimization method. *Appl Sci* 2022;12:6008. <https://doi.org/10.3390/app12126008>
- [35] Office for National Statistics. International territorial level; 2024 <https://www.ons.gov.uk/methodology/geography/ukgeographies/eurostat> [Accessed: 27 August 2025].
- [36] Office for National Statistics. Local authority district (April 2021) to LAU1 to ITL3 to ITL2 to ITL1 (January 2021) lookup in the United Kingdom, ONS Open Geography Portal, 2021. Open Government Licence v3.0. <https://geoportal.statistics.gov.uk/datasets/local-authority-district-april-2021-to-lau1-to-itl3-to-itl2-to-itl1-january-2021-lookup-in-the-uk> [Accessed: 15 July 2026].
- [37] Eurostat – GISCO. NUTS – territorial units for statistics: boundary dataset; 2024. NUTS 2021 version, 1:10M scale, EPSG:3035, GeoPackage format. <https://ec.europa.eu/eurostat/web/gisco/geodata/statistical-units/territorial-units-statistics> [Accessed: 13 March 2026].
- [38] Zhou Y, Essayeh C, Morstyn T. Datasets of Great Britain primary substations integrated with household heating information. *Data Br* 2024;54:110483. <https://doi.org/10.1016/j.dib.2024.110483>
- [39] Zhou Y, Essayeh C, Morstyn T. HDRah-Data-PS-GB: datasets of Great Britain primary substations, Zenodo, 2024. [Accessed: 15 July 2026]. <https://doi.org/10.5281/zenodo.10736516>, 4, 176 primary substations, 2021–22 reporting year; includes ONS LAD 2022 boundary files.
- [40] regiocom SE. GeoServer spatial data service — landkreis börde electrical grid infrastructure, internal GeoServer WFS, 2024. <https://webapp.regiocom.net/geoserver/web/?wicket:bookmarkablePage=:org.geoserver.web.demo.MapPreviewPage>. Layers used: v_epa_uw (substation locations), v_epa_uw_last (substation peak loads), v_lkb_vg250_gem_sql (Gemeinde boundaries), v_lkb_vg250_krs_sql (Kreis boundary). Accessed Mar 2026. Endpoint publicly reachable as of 2026-07-15 but not a permanent archive; Gemeinde/Kreis layers redistribute BKG VG250 (©GeoBasis-DE/BKG, dl-de/by-2-0).
- [41] OpenStreetMap contributors. OpenStreetMap land-use polygons; 2025. <https://www.openstreetmap.org/copyright>, landuse=* tags retrieved via the Overpass/OSMnx API; Open Database License (ODbL) 1.0. Accessed: 2025.
- [42] Earth Observation Group (EOG), NOAA/NCEI. VIIRS Day/Night band nighttime lights monthly composites, version 1; 2017. https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMISLFCG, Collection ID: NOAA/VIIRS/DNB/MONTHLY_V1/VCMISLFCG. Accessed via Google Earth Engine, median composite Jan 2021–Dec 2022.
- [43] Office for National Statistics. Regional gross value added (balanced) by industry: all ITL regions; 2024 <https://www.ons.gov.uk/economy/grossvalueadded/gva> [Accessed: 4 April 2025].
- [44] Eurostat. Gross value added at basic prices by NUTS 3 regions; 2026. https://doi.org/10.2908/NAMA_10R_3GVA, Online data code: nama_10r_3gva; values retrieved via the SDMX 2.1 dissemination API (unit CP_MEUR). https://ec.europa.eu/eurostat/databrowser/view/nama_10r_3gva/default/table?lang=en [Accessed: 13 March 2026].
- [45] Office for National Statistics. Estimates of the population for England and Wales; 2024 <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates> [Accessed: 4 April 2025].
- [46] Australian Bureau of Statistics. Australian statistical geography standard (ASGS) edition 3: digital boundary files (SA2/SA3/SA4, GDA2020); 2021. CC BY 4.0. <https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs/edition-3-july-2021-june-2026/access-and-downloads/digital-boundary-files> [Accessed: 14 July 2026].
- [47] Ausgrid. Distribution zone substation data; 2024. 15-min zone substation load, financial years 2008–2010 and 2024. Published under National Electricity Rules clause 5.13A; copyright retained by Ausgrid; catalogued on Data.NSW as “Distribution Zone Substation Information” under CC BY, <https://www.ausgrid.com.au/about-us/about-ausgrid/research-data-sets/distribution-zone-substation-data> [Accessed: 14 July 2026]. <https://data.nsw.gov.au/data/dataset/distribution-zone-substation>
- [48] NSW Spatial Services. Ausgrid UHC data (zone substation locations), ArcGIS hosted feature service; 2024. Extract date 2024-06-27; publicly accessible; “Data Sharing Agreement in place” terms apply. https://portal.data.nsw.gov.au/arcgis/rest/services/Hosted/Ausgrid_UHC_Data/FeatureServer/0 [Accessed: 14 July 2026].
- [49] Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES). Catchment scale land use of Australia (CLUM), 50 m raster: update March 2015 and update December 2023 (v2); 2023. <https://doi.org/10.25814/2w2p-ph98>, 2015-03 release (primary, NSW mapping ≈2007) via <https://data.gov.au/data/dataset/catchment-scale-land-use-of-australia-update-march-2015> [Accessed: 14 July 2026]. CC BY; 2023-12 v2 CC BY 4.0.
- [50] Earth Observation Group, NOAA/NCEI. DMSP-OLS nighttime lights time series, version 4; 2010. https://developers.google.com/earth-engine/datasets/catalog/NOAA_DMSP-OLS_NIGHTTIME_LIGHTS, Collection ID: NOAA/DMSP-OLS/NIGHTTIME_LIGHTS. Accessed via Google Earth Engine; median composite of F162008 and F162009. Image and data processing by NOAA’s National Geophysical Data Center; DMSP data collected by US Air Force Weather Agency.
- [51] Australian Bureau of Statistics. Regional population: population estimates by SA2 and above, 2001 to 2025; 2025. ERP backcast on ASGS Ed. 3 (2021) boundaries; CC BY 4.0. <https://www.abs.gov.au/statistics/people/population/regional-population/2024-25> [Accessed: 14 July 2026].
- [52] Australian Bureau of Statistics. Estimates of personal income for small areas (cat. No. 6524.0.55.002), 2005–06 to 2010–11; personal income in Australia, 2018–19 to 2022–23; 2025. CC BY 4.0. <https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/personal-income-australia> [Accessed: 14 July 2026].
- [53] Australian Bureau of Statistics. Census of population and housing 2011: basic community profile DataPack, SA3, New South Wales; 2012. Table B43, industry of employment; CC BY 2.5 AU. <https://www.abs.gov.au/census/find-census-data/datasets> [Accessed: 14 July 2026].
- [54] Thurner L, Scheidler A, Schäfer F, Menke J, Dollichon J, Meier F, Meinecke S, Braun M. pandapower — an open-source Python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Trans Power Syst* 2018;33:6510–21. <https://doi.org/10.1109/TPWRS.2018.2829021>
- [55] Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p -value calculation under arbitrary dependency structures. *J Am Stat Assoc* 2020;115:393–402. <https://doi.org/10.1080/01621459.2018.1554485>