forecasting

MDPI

Article

From Adoption Diffusion to Dimensioning: Probabilistic Forecasting of 5G/NB-IoT Demand via Monte Carlo Uncertainty Propagation

Nikolaos Kanellos [1], Dimitrios Katsianis [2,*] and Dimitris Varoutas [1]

[1] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimiopolis, T.Y.Π.A. Blgs. Ilisia, 15784 Athens, Greece; kanetza@di.uoa.gr (N.K.); d.varoutas@di.uoa.gr (D.V.)

[2] Department of Digital Industry Technologies, National and Kapodistrian University of Athens, NKUA Blgs., 34400 Psahna, Greece

* Correspondence: dkats@dind.uoa.gr

Highlights

What are the main findings?

- A diffusion-to-dimensioning workflow is proposed in which S-curve adoption modeling, bounded per-device usage priors, scenario stress testing, and Monte Carlo uncertainty propagation are combined to generate predictive 5G/NB-IoT demand distributions, exceedance curves, and quantile-based capacity rules.
- In the Great Britain case study (2025–2029), planning-year upper-tail risk is shown to be dominated by the asset-tracking class. Rank-based sensitivity indicates that AT payload size, burst factor, and message rate are the principal drivers of output variability, while proxy-scale perturbation and dependence–robustness tests show that upper-tail risk is driven more strongly by AT usage uncertainty than by proxy scale alone and is further amplified by positive AT co-movement.

What are the implications of the main findings?

- Medium-term 5G IoT planning should be shifted from deterministic point forecasting toward risk-based dimensioning because capacity headroom is set by upper-tail conditions rather than by average demand.
- Efforts to reduce uncertainty and over- or under-provisioning should be concentrated on the empirical characterization and operational governance of AT usage, payload, and burst behavior, since these inputs exert the strongest influence on capacity-relevant planning outcomes.

Academic Editor: Sonia Leva

Received: 29 January 2026
Revised: 18 March 2026
Accepted: 22 March 2026
Published: 25 March 2026

Abstract

Medium-term 5G/NB-IoT planning is made difficult by simultaneous uncertainty in device adoption and per-device traffic behavior because deterministic point forecasts do not quantify overload risk or support reliability-based capacity decisions. A diffusion-to-dimensioning workflow is proposed in which S-curve adoption modeling, bounded usage priors, scenario stress testing, and Monte Carlo uncertainty propagation are combined to generate predictive demand distributions, exceedance curves, and quantile-based capacity rules. The framework is applied to a Great Britain case study for 2025–2029 using smart meter deployment data and an M2M-based proxy for asset-tracking adoption. Analysis shows that planning-year upper-tail outcomes are driven primarily by asset-tracking usage uncertainty rather than by proxy scale alone. A ±30% perturbation of the AT adoption anchor changes $Q_{0.95}$ by approximately ±29.8%, whereas stressed AT usage increases $Q_{0.95}$

 https://doi.org/10.3390/forecast8020028

by 74.4%. Plausible positive dependence among key AT operational inputs further raises $Q_{0.95}$ by 18.3–22.5%. Limited hold-out evaluation provides strong out-of-sample support for the smart meter adoption stage and plausibility-only support for the shorter AT proxy. The framework is intended for medium-term, data-lean planning settings and is designed to support transparent risk-based capacity decisions rather than deterministic point sizing.

Keywords: 5G; IoT traffic forecasting; diffusion modeling; Monte Carlo simulation; capacity dimensioning

1. Introduction

Fifth generation (5G) mobile systems are expected to support not only enhanced mobile broadband but also massive machine-type communications (mMTCs), where large populations of low-complexity devices generate uplink-dominated traffic and signaling load [1,2]. For operators and regulators, medium-term planning for 5G IoT connectivity, therefore, requires demand models that remain credible when the installed base is still evolving and when uncertainty drivers are not only technological (e.g., coverage) but also market- and behavior-driven (e.g., device adoption and application usage intensity).

In IoT terms, the forecast target is shaped by heterogeneous device behaviors and operating modes across verticals [3]. Within the cellular ecosystem, machine-type communications have long been recognized as a distinct planning case, with differences in access patterns and load composition relative to human-centric traffic [4]. Large-scale measurements and traffic modeling studies further show that M2M/IoT traffic can deviate substantially from smooth “average-rate” assumptions, exhibiting heterogeneity and bursts across device types and applications [5,6]. This diversity motivates separating adoption uncertainty (how many devices) from usage uncertainty (how intensively they communicate) when producing demand forecasts for planning.

On the adoption side, S-shaped diffusion models, most prominently the Bass model and widely used logistic/Gompertz variants, provide a standard way to represent lifecycle growth under limited explanatory covariates [7–9]. On the engineering side, translating forecast traffic volumes into capacity decisions commonly relies on explicit traffic engineering definitions and scaling conventions (e.g., mean-rate equivalence or busy-hour scaling) [10]. However, deterministic point forecasts alone cannot answer risk-based planning questions, such as what capacity is sufficient so that overload remains rare under plausible futures.

This paper links diffusion and usage into a diffusion-to-dimensioning workflow that produces probabilistic demand forecasts and decision-facing risk outputs. Uncertainty is propagated using Monte Carlo simulation [11,12], enabling predictive distributions, exceedance probabilities, and quantile-based capacity rules. Where additional risk summaries are useful, tail-focused metrics, such as conditional value-at-risk (CVaR), may be used as optional extensions [13]. Uncertainty bounds and stress regimes are specified in a form consistent with standard quantitative risk analysis practice, emphasizing auditable min/mode/max elicitation when data are limited [14,15].

Methodologically, diffusion family selection is treated as an empirical modeling choice guided by fit diagnostics and information criteria (AIC) [16]. Where diffusion uncertainty is explicitly represented, adoption trajectories may be perturbed using an Itô process formulation around the fitted drift to generate ensembles of plausible diffusion paths [17,18], with innovation normality assessed using standard normality tests where relevant (e.g., Shapiro–Wilk) [19]. To support interpretable stress testing, sensitivity analysis may be included using rank-based screening (Spearman) and variance-based indices (Sobol/Saltelli) [20–22]. The implementation layer uses a practical Monte Carlo engine for generating risk outputs and sensitivity summaries.

Recent work on mobile and IoT traffic modeling has reinforced the need for demand representations that preserve heterogeneity across services, devices, and operating regimes. At the network level, recent traffic forecasting studies have shown that fine-grained dynamics, peak behaviour, and non-stationarity matter directly for resource allocation and medium- to long-term planning [23–25]. At the IoT access level, NB-IoT studies have likewise shown that application classes can differ materially in device behaviour, payload-related characteristics, and operating conditions, making a single homogeneous traffic assumption difficult to justify across verticals [26]. However, relatively limited work has combined medium-term adoption diffusion, class-specific usage uncertainty, and decision-facing exceedance-based capacity outputs within a single data-lean planning framework. That gap is addressed here by coupling proxy-based adoption diffusion with bounded per-device usage priors and Monte Carlo uncertainty propagation.

This paper makes four contributions. First, a telecom-oriented diffusion-to-dimensioning workflow is developed through which uncertain IoT adoption and per-device behavior are converted into predictive demand distributions and exceedance curves for medium-term planning. Second, adoption uncertainty and usage uncertainty are separated in a form that remains auditable in data-sparse settings through proxy-based diffusion fits, bounded usage priors, and structured stress regimes. Third, forecast uncertainty is translated into decision-facing outputs, including quantile-based capacity rules and exceedance probabilities, rather than being reported only as point estimates or interval bands. Fourth, through the GB case study, it is demonstrated how proxy-scale uncertainty, usage uncertainty, and operational dependence affect planning-year tail risk, with asset-tracking usage assumptions emerging as the dominant driver of upper-tail capacity requirements.

The contribution of the manuscript is forecasting- and risk-dimensioning-oriented rather than optimization-oriented. In particular, the proposed framework is designed to produce probabilistic demand distributions, exceedance curves, and quantile-based planning rules under uncertain adoption and usage; it is not formulated as a scenario-based MILP/MINLP infrastructure optimization model with explicit decision variables, a network cost objective, and solver-based optimality claims.

The framework is intended primarily for medium-term capacity planning in settings where both adoption and per-device behavior remain uncertain and where historical traffic measurements are sparse, short, or only partially representative. It is, therefore, most suitable for early-stage or data-lean planning environments in which transparent uncertainty propagation is needed before detailed network optimization becomes feasible.

The remainder of the paper is organized as follows: Section 2 presents the diffusion-to-traffic model, uncertainty specification, and risk outputs; Section 3 reports the GB case-study setup and results, including scenario outcomes and the sensitivity analysis; Section 4 discusses implications, limitations, and extensions; and Section 5 concludes.

2. Materials and Methods

Medium-term planning for 5G/NB-IoT is complicated by simultaneous uncertainty in adoption, per-device behavior, and the limited availability of long and representative service-specific traffic histories. To address this problem, a diffusion-to-dimensioning framework is formulated through which technology adoption diffusion is linked to per-device traffic usage and then propagated into probabilistic demand and risk outputs.

The framework is organized in five layers: diffusion modeling, traffic modeling, uncertainty specification, risk output generation, and, where required, sensitivity analysis.

Adoption trajectories are estimated by fitting S-curve families to adoption proxies. Aggregate traffic demand is then constructed from class-level adoption and per-device usage priors. Uncertainty in adoption and usage is propagated through Monte Carlo simulation, after which decision-facing outputs are derived in the form of predictive distributions, exceedance curves, and quantile-based capacity rules. When interpretive prioritization is needed, the uncertain inputs driving forecast variability are screened through global sensitivity measures.

For the reader's convenience, the principal symbols used in Equations (1)–(22) are summarized in Table 1.

Table 1. Notation.

Symbol	Meaning	Units	Domain/Status	First Used
t	month index	-	forecast month	Section 2.1
t_0	forecast origin	-	start of forecasting exercise	Section 2.1
H	forecast horizon	months or years	planning horizon	Section 2.1/Section 3
T	terminal month index/horizon length in months	-	upper limit in $t \in \{1, \ldots, T\}$	Section 2.1
$\mathcal{T}$	constrained month set/planning set	-	subset of months under planning constraint	Equation (1)
c	IoT class index	-	$c \in \{SM, AT\}$	Section System Model Overview/Equation (4)
D_t	aggregate traffic demand at month t	bits/month or Mbit/s	forecast target	Section 2.1
S_t	Signaling load proxy at month t	events/month or events/s	optional extension	Section 2.1
C, C_t	available capacity (static or time-varying)	same units as D_t	planning decision	Equation (1)
$T_{c,t}$	aggregate traffic generated by class c at month t	bits/month or Mbit/s	class-level traffic output	Equation (4)
$U_{c,t}$	average traffic generated per device	bits/device/month or equivalent rate	per-device traffic term	Equation (4)
${T_{c,t}}^{BH}$	busy-hour traffic for class c at month t	same units as $T_{c,t}$	peak-period engineering quantity	Equation (5)
κ_{BH}	busy-hour scaling factor/peak-to-average ratio	dimensionless	engineering mapping factor	Equation (5)
$A_{c,t}$	cumulative adoption/active subscription base for class c at month t	devices	diffusion-state variable	Equations (4), (7), and (8)
$N_{c,t}$	class-level device count entering the traffic model	devices	adopted class mapping $N_{c,t} = A_{c,t}$	Equation (6)
m_c	market potential/saturation anchor for class c	devices	diffusion input	Equation (7)
M	selected S-curve family	-	M ∈ {Bass, Logistic, Gompertz}	Equation (7)
$f_M(t; \theta_c, m_c)$	selected S-curve diffusion function for class c	devices	diffusion family representation	Equation (7)
θ_c	diffusion parameters for class c	model-specific	fitted parameters	Equation (7)
σ, σ_c	diffusion–volatility parameter	model-specific	generic in Equation (8); class-specific in implementation	Equations (8) and (19)
$\Delta W_t, dW_t$	Wiener increment/Brownian shock term	$\sqrt{time}$	stochastic diffusion term	Equations (8) and (19)
$r_{c,t}$	message rate	messages/device/month	usage input	Section 2.2/Equation (9)
$s_{c,}$	payload size	bits/message	usage input	Section 2.2/Equation (9)
$a_{c,t}$	active fraction/carried-share factor	0–1	bounded usage input	Section 2.2/Equation (9)

Table 1. *Cont.*

Symbol	Meaning	Units	Domain/Status	First Used
$b_{c,t}$	burst factor	dimensionless	usage input	Section 2.2/Equation (9)
$e_{c,t}$	signaling events per device	events/device/month	optional usage input	Section 2.2/Equation (10)
K	Monte Carlo sample size	trajectories	implementation setting	Section 2.3
$Q_q(D_t)$	q-quantile of predictive demand	same units as D_t	risk output	Equation (3), formalized in Equation (14)
$\Phi_t(C)$	exceedance curve at month t	probability	overload-risk function	Equation (16)
λ	cost/overload trade-off weight	0–1	optional decision extension	Equation (17)
Δ	average-month conversion constant	seconds/month	bits/month-to-rate conversion	Equation (20)

2.1. Problem Definition

For medium-term planning for 5G/NB-IoT, future demand depends on (a) diffusion/adoption of IoT devices and (b) per-device behavior that may be stable (e.g., smart metering) or bursty and mobility-driven (e.g., asset-tracking). The objective is to forecast predictive distributions (not only point forecasts) of aggregate demand in order to support risk-constrained dimensioning.

Let $t \in \{1, \ldots, T\}$ index months. At forecast origin t_0, probabilistic forecasts are produced over the required time span. The primary forecast target is the aggregate traffic demand D_t. To keep units explicit and implementation-agnostic, D_t may be expressed as:

- Traffic volume per month (e.g., bits/month);
- Equivalent mean rate (e.g., bits/s or Mbit/s), obtained by dividing monthly volume by the number of seconds in month t. If peak-period dimensioning is required, an additional busy-hour scaling factor may be applied.

Optionally, a signaling load proxy S_t may be forecast (e.g., events/month or events/s) to represent control-plane intensity associated with mobility, attachment, or session procedures.

Capacity planning depends on tail outcomes under uncertainty in both adoption and usage. A point forecast $\hat{D}_t$ does not quantify (i) the probability of exceeding a candidate capacity, (ii) the expected magnitude of overload, or (iii) the safety margin required to meet a reliability target. Predictive distributions are, therefore, required for D_t (and, when applicable, S_t) at each future month.

The planning decision is to choose a capacity C (static) or C_t (time-varying) such that overload risk is:

$$\mathbb{P}(D_t > C) \leq \alpha, \forall t \in \mathcal{T} \tag{1}$$

where $\alpha \in (0, 1)$ is the acceptable risk level and $\mathcal{T}$ denotes the set of months under the planning constraint.

Overload severity is quantified via expected exceedance:

$$\mathbb{E}[\max(0, D_t - C)] \tag{2}$$

which distinguishes between rare, mild violations and rare, severe violations.

A practical safety margin rule sets capacity to a high predictive quantile:

$$C_t = Q_{1-\alpha}(D_t) \tag{3}$$

so that, under the forecast distribution, $\mathbb{P}(D_t > C_t) \approx \alpha$.

System Model Overview

We define total aggregate network traffic demand as the sum of contributions from multiple IoT device classes (e.g., smart meters, asset trackers). For each class c at time t, the aggregate traffic $T_{c,t}$ is modeled as:

$$T_{c,t} = A_{c,t} \cdot U_{c,t} \tag{4}$$

where $A_{c,t}$ is the number of adopted/active devices and $U_{c,t}$ is the average traffic generated per device.

Traffic can be converted from an average monthly mean-rate equivalent into a busy-hour requirement using a busy-hour scaling factor κ_{BH}, interpreted in standard traffic-engineering terms as a peak-to-average ratio [10]. The busy-hour traffic $T_{c,t}^{BH}$ is then:

$$T_{c,t}^{BH} = \kappa_{BH} \cdot T_{c,t} \tag{5}$$

where $\kappa_{BH} \geq 1$.

2.2. Demand Modeling

A single planning region is considered (e.g., a market, city, or operator footprint) for which aggregate IoT-driven demand is forecasted at a monthly resolution. Spatial heterogeneity across cells and detailed radio-access scheduling effects are not resolved; the output is interpreted as offered aggregate load, to be mapped to capacity using an engineering rule (e.g., mean-rate equivalence or busy-hour scaling).

Two device classes are modeled: smart metering (SM) and asset-tracking/logistics (AT). The class index is $c \in \{\text{SM}, \text{AT}\}$.

Let $A_{c,t}$ denote cumulative adoption (or active subscriptions/active base) of class c at month t. In the adopted class-specific mapping, class-level device counts are defined as:

$$N_{c,t} = A_{c,t}, \; c \in \{SM, AT\} \tag{6}$$

Let $m_c > 0$ denote the market potential (saturation anchor) for class c. Diffusion dynamics for $A_{c,t}$ are represented using an S-curve family $M \in \{\text{Bass, Logistic, Gompertz}\}$, commonly used in innovation diffusion forecasting [7–9], with class-specific parameters θ_c:

$$A_{c,t} = f_M(t; \theta_c, m_c) \tag{7}$$

Model selection is based on fit diagnostics and information criteria (AIC) [16]. The fitted diffusion curve provides the expected adoption trajectory $E[A_{c,t}]$.

Furthermore, to represent residual diffusion uncertainty around the fitted adoption trajectory, we optionally introduce a stochastic diffusion formulation using an Itô process [17,18]. Let $A_{c,t}$ follow:

$$dA_{c,t} = \mu(A_{c,t}, t)\, dt + \sigma(A_{c,t}, t)\, dW_t \tag{8}$$

where dW_t is a Wiener increment and σ controls volatility [17]. Following this approach, the deterministic fitted diffusion curve is treated as the drift component, and stochastic perturbations calibrated from residual innovations [18] are introduced.

Residual innovations are computed from the fitted diffusion and tested for normality using the Shapiro–Wilk test [19]. Volatility σ is scenario-controlled using bounded perturbations (e.g., $\pm 10\%$) represented via a triangular distribution, consistent with common risk analysis elicitation practice when only plausible bounds are known [14,15].

Per-device behavior is represented using random variables that capture activity, message generation, payload size, and burstiness. For each class c and month t, the following quantities are defined:

- $r_{c,t}$: message rate (messages/device/month);
- s_c: payload size (bits/message);
- $a_{c,t}$: active fraction (dimensionless, $a_{c,t} \in [0,1]$); in the present interpretation, $a_{c,t}$ may also represent the fraction of class traffic carried by the dimensioned network (e.g., to represent "not all SM traffic is on cellular/5G" without changing notation);
- $b_{c,t}$: burst factor (dimensionless, $b_{c,t} \geq 1$).

An optional signaling characterization uses:

- $e_{c,t}$: signaling events per device per month (events/device/month).

Aggregate traffic demand D_t is constructed from class counts and per-device usage:

$$D_t = \sum\nolimits_c N_{c,t} \cdot a_{c,t} \cdot r_{c,t} \cdot s_c \cdot b_{c,t} \tag{9}$$

In volume form, D_t is expressed in bits/month. In mean-rate form, D_t may be divided by the number of seconds in month t and reported in bits/s (or Mbit/s). A busy-hour equivalence factor may be applied when peak-period dimensioning is required.

When signaling is modeled, the signaling load proxy is defined as:

$$S_t = \sum\nolimits_c N_{c,t} \cdot a_{c,t} \cdot e_{c,t} \tag{10}$$

with S_t in events/month (or events/s after rate conversion).

In the present application, capacity is sized using aggregate user-plane throughput equivalents derived from monthly traffic volume (bits/month) and mean-rate conversion. Signaling- or control-plane dimensioning is not included in the main GB results and is retained only as a practitioner-facing extension because the relevant bottlenecks and KPIs are deployment-specific.

Furthermore, D_t should be interpreted as an aggregate offered user-plane load for the planning region, expressed as a mean-rate equivalent. It is, therefore, most directly relevant to high-level planning layers, such as aggregate RAN headroom assessment, transport/backhaul provisioning, or regional/core demand screening, rather than to site- or cell-level scheduling. To convert D_t into an engineering dimensioning input, an additional mapping step is required. First, the monthly mean-rate equivalent should be converted to a busy-hour offered load using κ_{BH}. Second, that load should be allocated to the relevant planning layer (for example, across cells, sites, or transport links) using deployment-specific assumptions. The present GB results should, therefore, be interpreted as aggregate planning requirements rather than as direct cell-level engineered KPIs.

2.3. Uncertainty Introduction and Quantification

Uncertainty is represented through a separation between slowly varying/structural inputs (treated as trajectory-level draws) and usage variability (treated as month- and class-level draws).

Slowly varying inputs include diffusion parameters θ_c, market potential m_c, and (when enabled) diffusion volatility parameters for stochastic adoption trajectories. Usage variability includes $r_{c,t}$, $a_{c,t}$, $b_{c,t}$, and (optionally) $e_{c,t}$; payload sizes s_c are treated as fixed or narrowly distributed, depending on the application assumptions.

In addition to scenario stress testing of saturation and diffusion speed, optional diffusion uncertainty is introduced by perturbing the fitted diffusion trajectory using an Itô process approach consistent with the adoption-uncertainty formulation in [18]. Drift follows the fitted S-curve's implied growth rate; volatility is calibrated from drift-removed residuals/innovations of the adoption proxy series. Normality of the relevant innovation series is assessed using a standard normality test (e.g., Shapiro–Wilk) and reported for each adoption proxy series used in the case study.

Since scenario regimes already stress diffusion through the saturation anchor m_c and diffusion-speed multipliers $\mu_{g,c}$ (and $\mu_{m,c}$ when applicable), the Itô uncertainty component is included by varying only the volatility parameter σ_c.

Adoption is interpreted as active subscriptions/active base; therefore, monotonicity constraints are not enforced. Physical bounds apply; non-negativity is enforced, while an optional upper cap at m_c is reported as a robustness check.

Predictive uncertainty is propagated by generating K Monte Carlo trajectories [11,12]. A trajectory $k \in \{1, \ldots, K\}$ represents one plausible realization of the future over the horizon H, producing sequences $\left\{A_{c,t}^{(k)}\right\}$, $\left\{D_t^{(k)}\right\}$, and, when used, $\left\{S_t^{(k)}\right\}$. For each month t, the empirical sample $\left\{D_t^{(k)}\right\}_{k=1}^{K}$ defines the predictive distribution of D_t, from which medians, prediction intervals (e.g., 80% and 95%), and tail quantiles (e.g., $Q_{0.95}(D_t)$) are computed.

Proposed Uncertainty Parameterization

Scenario assumptions are parameterized using either distribution families (e.g., log-normal, gamma, beta) or bounded triangular specifications $(\text{min}, \text{mode}, \text{max})$ to support auditable elicitation.

Market potential m is bounded using demographic/industry constraints (e.g., addressable endpoints) and deployment limits. Growth parameters are bound to match plausible adoption speed (e.g., time-to-midpoint and saturation tail behavior). When a short adoption history exists, bounds are narrowed by matching early growth rates.

Smart metering (SM) is represented as stable periodic reporting with low variability and $b_{SM,t} \approx 1$. Asset-tracking (AT) is represented as operationally driven and potentially bursty behavior with wider uncertainty ranges and heavier tails, particularly for $b_{AT,t}$. When a stress regime intends to represent technology/use case market failure without modifying adoption, the regime is implemented by lowering $a_{c,t}$ and/or $r_{c,t}$ while keeping $A_{c,t}$ unchanged.

Last, when monthly seasonality or operational cycles are expected, scenario definitions may include seasonal modifiers applied to $r_{c,t}$ or $a_{c,t}$ (e.g., multiplicative factors with mean 1). Absent supporting evidence, parameters are treated as stationary in time to reduce assumption complexity.

2.4. Risk Metrics and Dimensioning Rule

2.4.1. Exceedance Probability and Severity

Let C denote available capacity expressed in the same units as D_t (e.g., bits/month or mean rate). Overload risk at month t is quantified through the exceedance probability $\mathbb{P}(D_t > C)$.

Given Monte Carlo samples $\left\{D_t^{(k)}\right\}_{k=1}^{K}$, the exceedance probability is estimated by:

$$\hat{\mathbb{P}}(D_t > C) = \frac{1}{K}\sum\nolimits_{k=1}^{K} 1\{D_t^{(k)} > C\} \tag{11}$$

When signaling is modeled and signaling capacity $C^{(S)}$ is specified, an analogous exceedance probability $\mathbb{P}\left(S_t > C^{(S)}\right)$ is computed.

Exceedance probability does not distinguish between minor and severe overload. Severity is quantified using expected exceedance:

$$\mathbb{E}[\max(0, D_t - C)] \tag{12}$$

estimated from Monte Carlo samples as:

$$\hat{\mathbb{E}}[\max(0, D_t - C)] = \frac{1}{K}\sum\nolimits_{k=1}^{K} \max(0, D_t^{(k)} - C) \tag{13}$$

Expected exceedance supports trade-off analysis between capacity cost and expected overload magnitude.

2.4.2. Quantile-Based Safety Margin Rule

A quantile-based rule is adopted to translate predictive distributions into a capacity target under risk tolerance $\alpha \in (0,1)$. Capacity is set to the $(1-\alpha)$ quantile of the predictive distribution:

$$C_t = Q_{1-\alpha}(D_t) \tag{14}$$

Under the predictive distribution, this rule implies $\mathbb{P}(D_t > C_t) \approx \alpha$.

For a constant capacity decision over a set of constrained months $\mathcal{T} \subseteq \{t_0 + 1, \ldots, t_0 + H\}$, a conservative choice is:

$$C = \max_{t \in \mathcal{T}} Q_{1-\alpha}(D_t) \tag{15}$$

which controls exceedance risk across $\mathcal{T}$ under the predictive distribution.

2.4.3. Risk Profiles and Exceedance Curves

For a fixed month t, an exceedance curve summarizes overload risk across candidate capacities:

$$\phi_t(C) = \mathbb{P}(D_t > C) \tag{16}$$

The curve $\phi_t(C)$ is monotone decreasing in C and can be estimated from the empirical predictive distribution by evaluating $\hat{\mathbb{P}}(D_t > C)$ over a grid of C values. Exceedance curves support the selection of C for alternative risk tolerances and enable comparison of planning months (e.g., seasonal peaks).

2.4.4. Optional Decision Extensions

When a cost model for capacity is available, C may be selected by minimizing a weighted objective, such as:

$$\min_{C} \ \lambda \, \mathrm{Cost}(C) + (1-\lambda) \, \mathbb{E}[\max(0, D_t - C)] \tag{17}$$

with $\lambda \in [0,1]$. This extension is not required for risk-based planning but provides an additional interpretation of expected exceedance as an overload penalty.

2.5. Predictive Checks and Robustness Evaluation

Evaluation is performed primarily through structured scenario analysis because reliable out-of-sample demand history for the target 5G/NB-IoT service classes is not available. A scenario set is defined by (i) diffusion/adoption assumptions and (ii) a Monte Carlo sampling plan that generates predictive samples $\left\{D_t^{(k)}\right\}_{k=1}^{K}$ (and optional $\left\{S_t^{(k)}\right\}$) over the horizon.

However, an empirical hold-out check is also conducted for the observable adoption stage of the framework using the available real proxy series. In this way, the diffusion stage is examined against later observed proxy values, which provides a direct empirical check on growth realism and out-of-sample stability where observable data exist. Because independent measured 5G/NB-IoT service traffic series were not available for the target classes, this exercise should be interpreted as validation of the proxy-adoption stage rather than as end-to-end validation of the final Monte Carlo demand forecasts.

Decision-relevant outputs are extracted from predictive samples for each scenario set, including:

$$Q_{1-\alpha}(D_t), \mathbb{P}(D_t > C), \mathbb{E}[\max(0, D_t - C)] \quad (18)$$

Scenario-based evaluation is reported using: (i) fan charts of D_t (median with 80/95% intervals), (ii) tail quantiles (e.g., $Q_{0.95}(D_t)$), (iii) exceedance curves $\phi_t(C)$ for selected months, and (iv) capacity trajectories $C_t = Q_{1-\alpha}(D_t)$ under nominal and stress regimes.

2.5.1. Robustness and Stress Testing

Robustness is assessed by comparing predictive outputs across a small number of structured stress regimes. A minimal set includes:

- Nominal scenario: central parameter values (e.g., modes/medians) with baseline variability.
- High-adoption stress: upper-range m and higher diffusion growth (e.g., higher q or g).
- High-usage stress: upper-range $r_{c,t}$ and payload s_c, with increased activity $a_{c,t}$.
- High-burst stress (AT-focused): upper-range $b_{AT,t}$ and optional signaling intensity $e_{AT,t}$.
- Compound stress: combined high-adoption, high-burst, and high-usage to test worst-case tail risk.

Robustness is reported via changes in tail and risk metrics at decision-relevant months. Capacity implications are summarized using $C_t = Q_{1-\alpha}(D_t)$ under each stress regime.

2.5.2. Baselines

Two baselines are included to contextualize the proposed uncertainty propagation.

- B1: Deterministic mechanistic baseline. Diffusion is fixed, and all usage inputs are fixed at nominal values, producing a point trajectory D_t. This baseline provides a reference for the magnitude of uncertainty bands and quantile-based safety margins.
- B2: Single-model probabilistic baseline. Diffusion model uncertainty is disabled, while parameter and usage variability are retained. Optional Itô process diffusion uncertainty is also disabled so that the incremental impact of stochastic diffusion perturbations on upper-tail outputs and on C can be isolated by comparison with the full method.

2.6. Sensitivity Analysis

Sensitivity analysis may be included to identify which uncertain inputs most influence decision-relevant planning outputs. The analysis is performed on scalar outputs derived from the predictive distribution at a target month t^* (e.g., $t^* = t_0 + 36$) or over a constrained month set $\mathcal{T}$. It should be noted that the deterministic baseline B1 fixes inputs at nominal point values (modes), whereas B2 summarizes the predictive distribution under the same nominal scenario. For the AT class, several priors are intentionally right-skewed, and the multiplicative demand construction amplifies this asymmetry. Consequently, the distribution median may exceed the mode-based baseline, and B1 should be interpreted as a most likely value reference rather than as a central estimate.

A two-stage procedure is recommended: (i) rank-based screening (e.g., Spearman [20]) to prioritize inputs, followed by (ii) variance-based Sobol indices [21,22] for the highest-ranked inputs when computational budget permits. The results are reported as a ranked table or tornado plot for screening and, when computed, a compact table of first-order and total-order Sobol indices for the top inputs.

Recommended outputs include: $Y_1 = Q_{0.95}(D_{t^*})$, $Y_2 = \mathbb{P}(D_{t^*} > C)$ and $Y_3 = \mathbb{E}[\max(0, D_{t^*} - C)]$.

3. Results

To showcase the proposed uncertainty-propagation framework, an application over Great Britain (GB) is presented. The case study converts uncertain IoT diffusion and per-device usage into predictive distributions of monthly aggregate demand D_t and corresponding risk-based planning outputs, including exceedance probabilities, exceedance curves $\phi_t(C) = \mathbb{P}(D_t > C)$, and quantile-based capacity trajectories $C_t = Q_{1-\alpha}(D_t)$ with default $\alpha = 0.05$.

The model is evaluated at a monthly resolution with month index t, forecast origin t_0, and forecast horizon H of 5 years. Two IoT classes are considered, $c \in \{\mathrm{SM}, \mathrm{AT}\}$, representing smart metering (SM) and asset-tracking/logistics (AT). Class-level device counts follow the mapping $N_{c,t} = A_{c,t}$, where $A_{c,t}$ denotes cumulative adoption (or active subscriptions/active base) for class c. Monthly aggregate demand is computed as the traffic volume D_t in bits/month using Equation (9).

The probabilistic assessment is reported for the planning year 2029 as the end-of-horizon design year, while intermediate years are contextualized using baseline trajectories (B1 and B2). Demand and capacity are reported primarily as mean-rate equivalents (Mbit/s), obtained by converting monthly volumes (bits/month) using the adopted month-to-seconds convention; the corresponding bits/month values are provided as reference.

Busy-hour scaling and expected exceedance $\mathbb{E}[\max(0, D_t - C)]$ are reported separately as supporting analysis rather than as primary results in Section 3.

3.1. Case Study Formulation and Inputs

3.1.1. Adoption Proxies and Market Potential Anchors

Smart metering adoption for GB is represented using DESNZ smart meter statistics tables [27], using domestic and non-domestic "total smart meters" to construct an SM adoption proxy. Total meter stock (smart and non-smart meters) provides a saturation (market potential) anchor for SM, consistent with the interpretation of m_{SM} as an upper bound on addressable devices.

Direct AT adoption series were not available at the national scale in a consistent form. Therefore, annual M2M mobile network subscriptions were used as an AT scale anchor, combining 2012–2015 values from Ofcom reporting [28] with 2016 onward values from the ITU "M2M mobile-network subscriptions" indicator [29,30]. The M2M series was not interpreted as a direct count of asset-tracking devices. Instead, it was used to provide a plausible scale for the AT adoption path. Because M2M subscriptions are not uniquely attributable to asset-tracking, wider AT uncertainty bounds were retained to reflect proxy mismatch and coverage ambiguity. In addition, a dedicated sensitivity test was performed, in which the baseline AT adoption path was perturbed multiplicatively ($0.7\times$, $1.0\times$, $1.3\times$) while AT usage priors were held fixed, and the resulting effect was compared with a separate stressed-usage case.

3.1.2. Time Resolution and Disaggregation

Adoption inputs for the GB application are reported at quarterly resolution for SM and annual resolution for the M2M proxy used to scale AT. The model remains monthly as defined in Section 2. Adoption is treated as piecewise-constant within the reporting period for months lacking finer resolution. Specifically, $N_{AT,t}$ is held constant within each calendar year, while $N_{SM,t}$ is held constant within each calendar quarter, unless interpolation is explicitly enabled.

Demand is computed as monthly volume (bits/month) because $r_{c,t}$ is specified in messages/device/month. Exceedance curves $\phi_t(C)$ and capacity trajectories C_t are computed on the monthly D_t distributions. When adoption is constant within a year (AT) or quarter (SM) and no seasonality is applied to usage variables, within-period months may share the same $\phi_t(C)$ and C_t; any remaining differences arise from sampled per-device usage variability.

Since adoption is treated as piecewise-constant within the reporting period and no seasonality is imposed by default, the probabilistic results reported for 2029 correspond to the distribution of a representative month in the planning year under the adopted conversion convention. The resulting exceedance curves and quantile-based capacity are, therefore, interpreted as monthly planning metrics for 2029. In addition, a dedicated robustness check of this within-period implementation choice was performed in the Section 3 by comparing the baseline piecewise-constant treatment with interpolated monthly adoption paths while keeping the Monte Carlo structure and all other model settings unchanged.

3.2. Parameterization and Scenario Implementation

3.2.1. Diffusion Fitting for SM and AT Adoption Proxies

Adoption trajectories $A_{c,t}$ are generated by fitting S-curve diffusion families to the GB adoption proxies described in Section 3.1.2. For smart metering (SM), the DESNZ proxy supports a stable and well-behaved diffusion fit, and the best-fitting family is selected using standard goodness-of-fit diagnostics (e.g., residual inspection and information criteria where applicable). For asset-tracking/logistics (AT), the M2M subscriptions proxy exhibited structural features not uniquely attributable to AT, and the fitted trajectory was, therefore, treated as a smooth approximation of an AT scale adoption path rather than as a direct measurement of AT diffusion. This proxy mismatch motivated wider uncertainty bounds for AT relative to SM in the scenario analysis and was further examined through a separate adoption anchor sensitivity test in Section 3. In addition, an asymmetric hold-out exercise was performed for the adoption stage; a substantive out-of-sample check was applied to the domestic and non-domestic SM proxy series, while only a limited plausibility check was applied to the shorter annual M2M proxy used for AT.

If diffusion uncertainty is enabled, stochastic adoption trajectories are generated using an Itô process perturbation around the fitted diffusion drift, consistent with Section 2.3. In this case, drift follows the fitted S-curve's implied growth rate, and volatility is calibrated from drift-removed residuals/innovations of the proxy series. Normality of the innovation series is assessed using a standard test (e.g., Shapiro–Wilk), and the tested series and test outcomes are reported for each proxy. For uncertainty propagation, the volatility parameter σ_c is sampled per trajectory from a triangular distribution $(0.9\sigma_c, \sigma_c, 1.1\sigma_c)$.

Although Equation (8) is written in continuous-time form, trajectories are generated on the monthly planning grid using an Euler–Maruyama discretization. For month index t with step Δt (in years), Brownian increments are simulated as $\Delta W_t \sim \mathcal{N}(0, \Delta t)$, yielding

$$A_{c,t+\Delta t} = A_{c,t} + \mu_c(A_{c,t}, t)\Delta t + \sigma_c A_{c,t} \Delta W_t \quad (19)$$

where $\mu_c(\cdot)$ follows the fitted S-curve drift and σ_c is calibrated from drift-removed proxy residuals; Sampling is performed per trajectory from Triangular$(0.9\sigma_c, \sigma_c, 1.1\sigma_c)$.

The resulting class-specific diffusion-family choices, fitted parameters, fit diagnostics, and diffusion-uncertainty settings are summarized in Table 2.

Table 2. Diffusion specification per class.

Class/Proxy	Selected Family	Fitted Parameters	Fit Diagnostics	Itô Process Diffusion Uncertainty	Shapiro–Wilk Test
Domestic SM	Gompertz	Saturation = 45,486,650 Rate r_d = 0.067725 Time t_d = 29.6006	n = 53 R2 = 0.999 RMSE = 497,616 MAE = 359,041	Calibrated σ = 0.002814 Triangular (0.9σ, σ, 1.1σ)	W = 0.8729 p-value = 0.0711
Non-domestic SM	Logistic	Saturation = 782,446 Rate r_d = 0.124457 Time t_d = 40.2256	n = 49 R2 = 0.998 RMSE = 7237 MAE = 5098	Calibrated σ = 0.007932 Triangular (0.9σ, σ, 1.1σ)	W = 0.8842 p-value = 0.0993
AT proxy (M2M subscriptions)	Logistic	Saturation = 109,237,867 Rate r_d = 0.20893 Time t_d = 40.2256	n = 13 R2 = 0.905 RMSE = 2,640,420 MAE = 2,249,424	Calibrated σ = 0.018426 Triangular (0.9σ, σ, 1.1σ)	W = 0.9634 p-value = 0.8056

Note: SM total adoption is implemented as the sum of domestic and non-domestic components; diffusion families and parameters are reported for each component used in the sum. Note 2: Normality test results are reported for the calibration window used for volatility-style checks (the most recent 12 quarters, i.e., Q4 2022–Q3 2025, for SM domestic and SM non-domestic). The AT proxy (M2M) is annual and uses the full observed window (2012–2024). When tested over the full SM observation window (Q3 2012–Q3 2025), residual normality is rejected (domestic: $n = 53$, $W = 0.9366$, $p = 0.0074$; non-domestic: $n = 49$, $W = 0.8759$, $p = 9.81 \times 10^{-5}$), indicating that departures from normality are driven mainly by early diffusion-phase deviations rather than the recent calibration window.

3.2.2. Per-Device Usage Priors and Constraints

Per-device usage variables are specified on a monthly basis and sampled independently unless otherwise stated. The GB application adopts auditable priors parameterized by min/mode/max triples (triangular distributions), with boundedness constraints enforced as required by variable support. In particular, the activity factor $a_{c,t} \in [0,1]$ is sampled from a bounded distribution (triangular truncated/capped in [0, 1], or beta if used), and scenario multipliers applied to $a_{c,t}$ are capped at 1. In addition to the independent baseline, an optional dependence–robustness variant was also evaluated. Positive dependence was imposed only among the AT message rate, AT burst factor, and AT activity fraction, while AT payload size and all other model inputs were left unchanged. The tested correlation levels were treated as illustrative robustness settings rather than as empirically estimated behavioral parameters.

The smart metering class is parameterized as stable periodic reporting with narrow uncertainty ranges and $b_{SM,t} \approx 1$. The asset-tracking class is parameterized with wider uncertainty ranges to reflect operational heterogeneity and tail risk, with burstiness captured through $b_{AT,t}$. Payload sizes s_c are defined in bits/message and treated as class-constant (or narrowly distributed) to maintain unit consistency in Equation (9).

The class-specific usage priors, units, distribution families, and support bounds adopted for the GB application are summarized in Table 3.

Signaling load is not modeled in the present application and is left as an implementation choice for practitioners.

Table 3. Usage prior distributions entered for the GB application ($r_{c,t}$, $a_{c,t}$, $b_{c,t}$, s_c), including units and bounds.

Class c	Variable	Meaning	Units	Distribution Family	Min	Mode	Max	Support/Bounds
SM	$r_{SM,t}$	message rate	messages/device/month	Triangular	120	150	180	>0
SM	$a_{SM,t}$	active fraction (includes “cellular-carried share”)	0–1	BetaPert	0.2	0.4	0.6	Capped to [0, 1] after scenario multipliers
SM	$b_{SM,t}$	burst factor	≥1	Triangular	1	1.05	1.1	≥1
SM	s_{SM}	payload size	bits/message	Triangular	12,800	16,000	19,200	>0
AT	$r_{AT,t}$	message rate	messages/device/month	Triangular	1440	3000	8640	>0
AT	$a_{AT,t}$	active fraction	0–1	BetaPert	0.6	0.85	1	Capped to [0, 1] after scenario multipliers
AT	$b_{AT,t}$	burst factor	≥1	Triangular	1.5	3	10	≥1
AT	s_{AT}	payload size	bits/message	Triangular	400	1600	6400	>0

3.2.3. Scenario Multipliers and Orthogonality of Stress Regimes

The GB application implements the stress test scenario regimes defined in Section 2.5.1: nominal, high-adoption, high-usage, high-burst (AT-focused), and compound. Adoption stress is applied through class-specific multipliers on market potential and diffusion speed $(\mu_{m,c}, \mu_{g,c})$, with stronger perturbations for AT to reflect the proxy mismatch and wider plausibility bounds. In the scenario matrix, $\mu_{m,c}$ scales the market potential m_c, and $\mu_{g,c}$ scales the diffusion speed parameter(s) of the selected S-curve family (e.g., the logistic/Gompertz rate term), leaving the functional form unchanged.

Usage stress is applied multiplicatively to the base prior distributions. The high-usage regime increases $r_{c,t}$, $a_{c,t}$, and s_c (with stronger multipliers for AT) while keeping $b_{c,t}$ at nominal levels. The high-burst regime targets AT tail behavior by increasing $b_{AT,t}$. To maintain orthogonality between regimes, $r_{AT,t}$ and $a_{AT,t}$ may hold near nominal values in the high-burst regime, so that tail changes are primarily attributable to burstiness rather than average message generation.

The compound regime combines high-adoption, high-usage, and high-burst multipliers. For bounded variables, $a_{c,t}$ is multiplied by the scenario factor and capped at 1.

The full set of scenario multipliers implemented for the GB application is summarized in Table 4.

Table 4. Scenario matrix implemented for the GB application.

Scenario	$\mu_{m,SM}$	$\mu_{m,AT}$	$\mu_{g,SM}$	$\mu_{g,AT}$	r_{SM}	r_{AT}	a_{SM}	a_{AT}	b_{SM}	b_{AT}	s_{SM}	s_{AT}	e_{AT} **(If Used)**
Nominal	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
High-adoption	1.05	1.25	1.10	1.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
High-usage	1.00	1.00	1.00	1.00	1.10	1.25	1.05	1.10	1.00	1.00	1.10	1.15	1.00
High-burst (AT)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.75	1.00	1.00	1.75
Compound	1.05	1.25	1.10	1.25	1.10	1.25	1.05	1.10	1.00	1.75	1.10	1.15	1.75

Note: Signaling is not applied in the main GB results; the e_{AT} column is retained only as a practitioner-facing extension.

3.2.4. Monte Carlo Implementation and Reported Outputs

Uncertainty is propagated using Monte Carlo sampling. The model generates K stochastic samples/trajectories for the planning year 2029. The empirical distribution $\{D_{2029}^{(k)}\}_{k=1}^{K}$ is computed from Equation (9) and summarized using the median and central prediction intervals (80% and 95%). Tail behavior relevant to planning is reported using upper quantiles (e.g., $Q_{0.95}(D_{2029})$).

Risk-based planning outputs are derived directly from the sampled distribution for 2029. The exceedance curve $\phi_{2029}(C) = \mathbb{P}(D_{2029} > C)$ is evaluated for a range of capacities C, and the quantile-based capacity $C = Q_{1-\alpha}(D_{2029})$ is reported for $\alpha = 0.05$. Demand and capacity are presented primarily in Mbit/s, with bits/month retained as a traceable reference. Busy-hour scaling and expected exceedance $\mathbb{E}[\max(0, D_{2029} - C)]$ are reported separately as supporting analysis.

Computational requirements were modest for the present GB application. The Monte Carlo engine generated $K = 10,000$ trajectories for the planning-year evaluation at monthly resolution. On the data side, the diffusion-fitting stage used 53 quarterly observations for domestic SM, 49 quarterly observations for non-domestic SM, and 13 annual observations for the AT/M2M proxy. In practical terms, externally anchored S-curve fitting is workable with approximately 10–15 observations, although shorter or lower-frequency series reduce the ability to discriminate among diffusion families and weaken any out-of-sample assessment.

The fitted diffusion trajectories for the GB smart-metering and asset-tracking adoption proxies, together with the corresponding residual diagnostics, are shown in Figure 1. Illustrative effects of the scenario multipliers on selected AT per-device usage priors are shown in Figure 2.

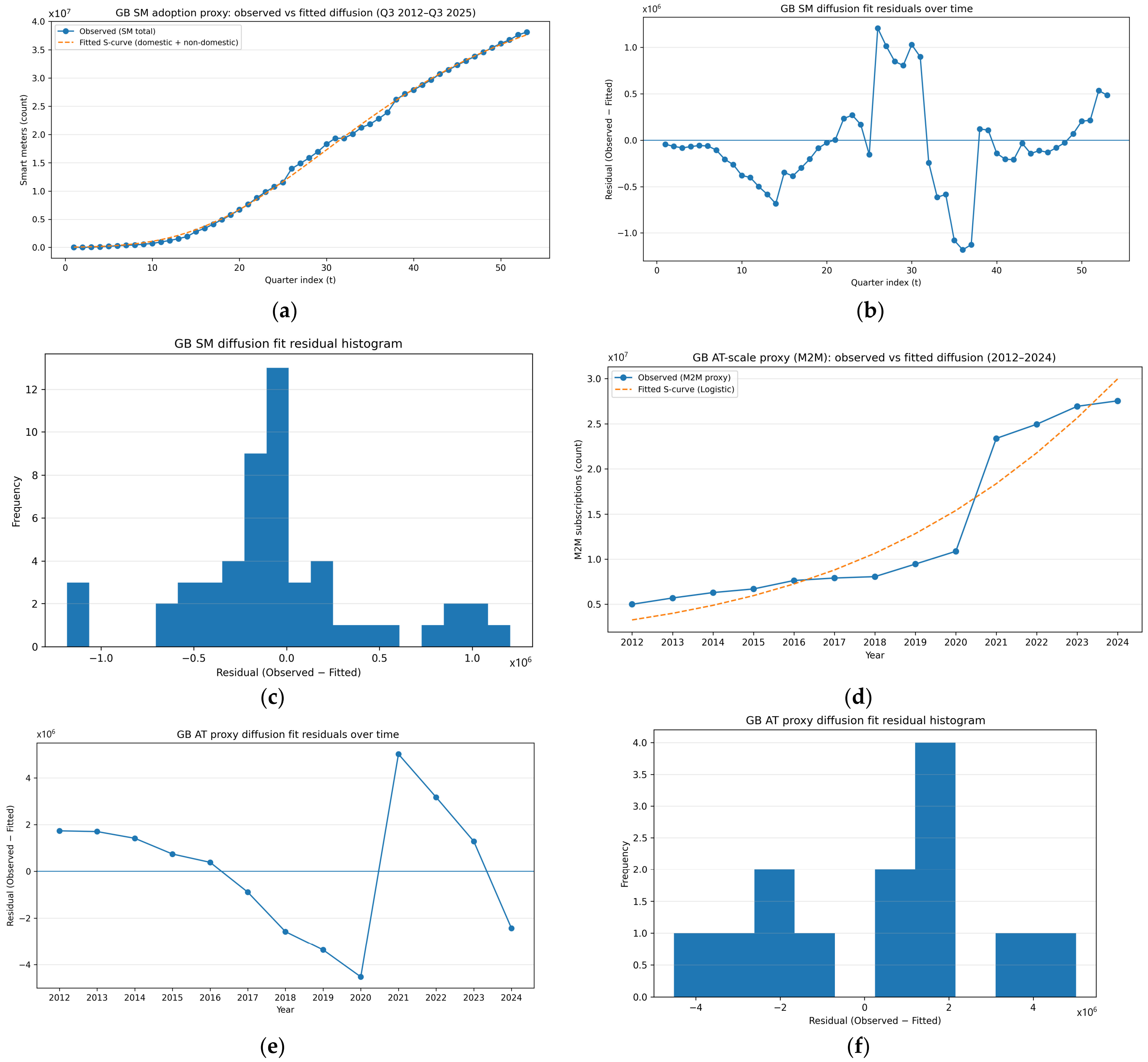
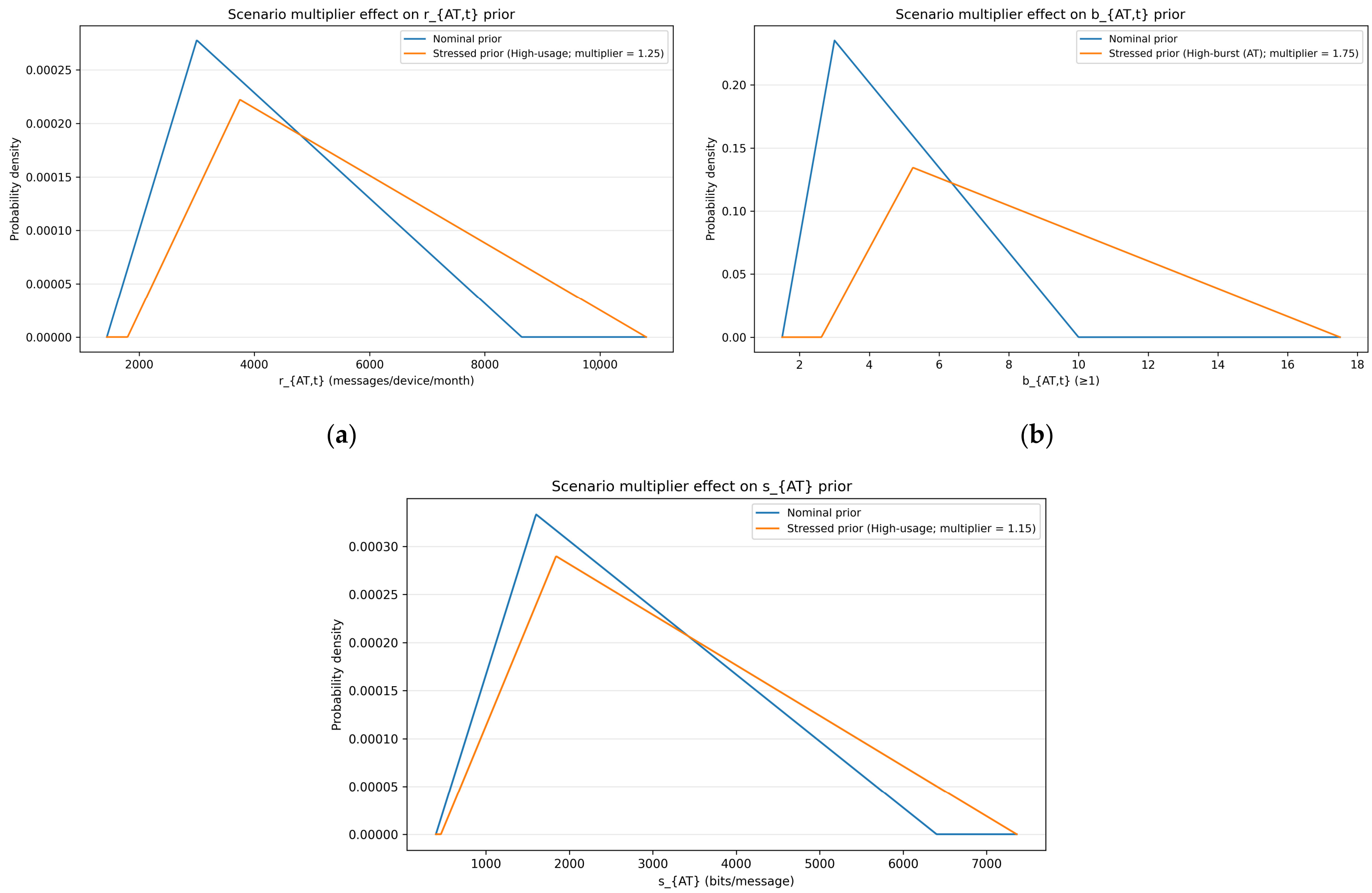


Figure 1. S-curve diffusion fits for GB adoption proxies and residual diagnostics. (**a**) Smart metering (SM) adoption proxy (DESNZ "total smart meters", domestic + non-domestic): observed quarterly series and fitted diffusion trajectory (sum of domestic and non-domestic fitted components). (**b**) SM fit residuals (observed–fitted) over the quarterly index. (**c**) Histogram of SM fit residuals. (**d**) Asset-tracking (AT) scale proxy: annual M2M mobile network subscriptions (Ofcom for 2012–2015; ITU from 2016 onward) with fitted logistic diffusion trajectory. (**e**) AT proxy fit residuals over time. (**f**) Histogram of AT proxy fit residuals. Residual diagnostics are included to illustrate goodness of fit and to motivate uncertainty bounds for the proxy-based AT diffusion.



(**a**) (**b**)

(**c**)

Figure 2. Illustration of scenario multiplier effects on selected per-device usage priors (nominal vs. stressed). (**a**) AT message rate $r_{AT,t}$: nominal triangular prior and the stressed prior obtained by applying the high-usage multiplier (as specified in the scenario matrix) to the min/mode/max triple. (**b**) AT burst factor $b_{AT,t}$: nominal triangular prior and the stressed prior under the high-burst (AT-focused) regime. (**c**) AT payload size s_{AT}: nominal triangular prior and the stressed prior under the high-usage regime. Stressed priors are produced by multiplicative scaling of the base priors, with bounded variables (when applicable) capped to their physical support as specified in Section 2.

3.3. Predictive Demand Results

The probabilistic demand results are reported for planning year 2029 under the scenario regimes defined in Section 2.5.1. Monte Carlo simulation generates $K = 10{,}000$ samples of the monthly aggregate demand D_{2029} for each scenario, from which empirical prediction intervals and tail quantiles are estimated.

Demand is computed in bits/month and reported primarily as a mean-rate equivalent. Using the average month convention $\Delta = 2{,}629{,}800$ s, we convert volume to rate as:

$$D_t^{(\mathrm{Mbit/s})} = \frac{D_t^{(\mathrm{bits/month})}}{\Delta \cdot 10^6} \tag{20}$$

3.3.1. Scenario-Level Distributions of Total Demand

Table 5 summarizes the predictive distribution of total demand in 2029 for each scenario using the median, the central 80% prediction interval (P10–P90), the central 95% prediction interval (P2.5–P97.5), and the 0.95 quantile $Q_{0.95}(D_{2029})$, which corresponds to the quantile-based capacity used for $\alpha = 0.05$. The results are reported in Mbit/s, with bits/month provided as a reference.

Table 5. Predictive summaries for total demand D_{2029} under each scenario.

Scenario	Median (Mbit/s)	P10–P90 (Mbit/s)	P2.5–P97.5 (Mbit/s)	$Q_{0.95}(D_{2029})$ (Mbit/s)
Nominal	750.5	287.8–1861	168.7–2824	2351.8
High-adoption	904.6	342.6–2227.9	199–3422.3	2846.4
High-usage	1180.3	446.7–2900.1	262.3–4474.5	3728.2
High-burst (AT)	1304.4	488.9–3215.5	282.9–4976.6	4098.8
Compound	2473.1	924.7–6131.1	531.1–9382	7830.3

The scenario-level predictive distribution of total demand in the 2029 planning year is visualized in Figure 3.

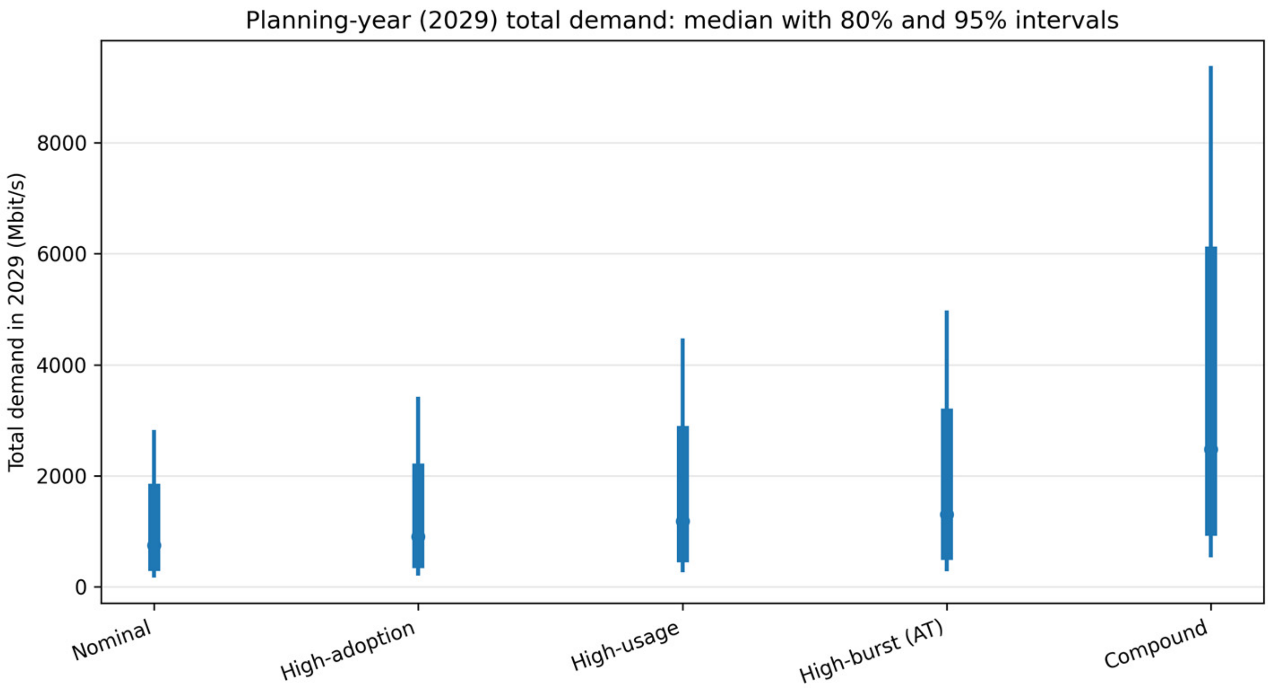


Figure 3. Planning-year (2029) predictive distribution of total demand D_{2029} across scenarios (median with 80% and 95% prediction intervals).

Across scenarios, median demand increases monotonically from nominal to compound. Tail behavior exhibits stronger amplification than the median, particularly in the high-burst and compound regimes, indicating that burstiness and usage uncertainty contribute materially to planning-relevant upper quantiles. The compound regime produces the largest right-tail shift, reflecting the combined action of adoption and usage stresses.

To clarify the class-level source of aggregate uncertainty, the 2029 per-class demand distributions under the nominal and stressed AT regimes are shown in Figure 4.

Asset tracking dominates the upper tail because the per-device demand model is multiplicative in adoption and usage components (e.g., $D_{c,t}$ scales with $A_{c,t}\, r_{c,t}\, a_{c,t}\, b_{c,t}\, s_c$, and related factors). Relative to smart metering (SM), AT is assigned a higher message intensity $r_{AT,t}$, a higher activity $a_{AT,t}$, and—most importantly—a substantially larger burst factor $b_{AT,t}$, which amplifies peak-equivalent demand under busy-hour conversion. In addition, the AT payload size s_{AT} is modeled with wider uncertainty than SM, so even when medians are comparable, AT contributes disproportionately to high quantiles (e.g., $Q_{0.95}$).

It should be noted that, although 7.8 Gbit/s may appear modest relative to broadband traffic volumes, the present application concerns fragmented, high-heterogeneity mMTC demand. Operational complexity may, therefore, remain high, even when aggregate throughput remains moderate.

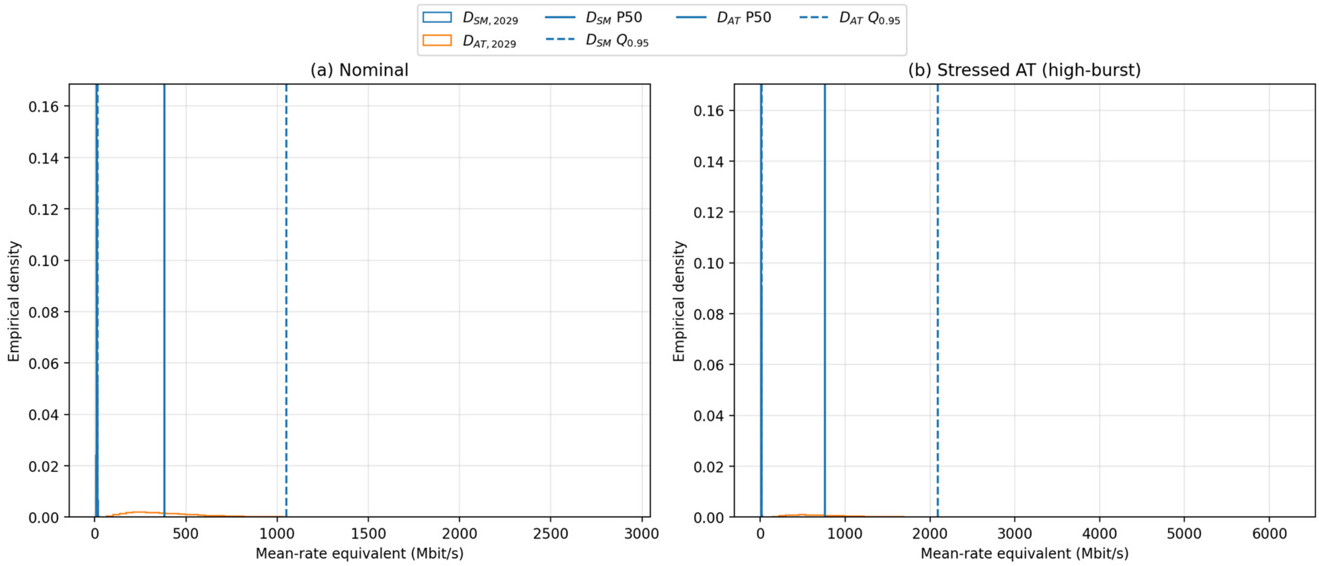


Figure 4. Per-class demand distributions (2029). Empirical distributions of $D_{SM,2029}$ and $D_{AT,2029}$ across $K = 10,000$ Monte Carlo runs under (**a**) the nominal scenario and (**b**) the stressed AT (high-burst) scenario. Vertical lines indicate P50 and $Q_{0.95}$ for each class. (Units shown as mean-rate equivalent, Mbit/s).

3.3.2. Class Contributions (SM vs. AT)

To interpret the aggregate distributions, class contributions are examined using the sampled components of Equation (9). Table 6 reports the median contributions of AT and total SM demand in Mbit/s and the distribution of the AT share D_{2029}^{AT}/D_{2029} across Monte Carlo trials. The decomposition of total 2029 demand into AT and SM components across scenarios is shown in Figure 5.

Table 6. 2029 class contribution summaries.

Scenario	AT Median (Mbit/s)	Total SM Median (Mbit/s)	AT Share (Median Across Trials)	AT Share (P10–P90 Across Trials)
Nominal	733.5	16.72	97.8%	94–99.1%
High-adoption	887.2	16.48	98.2%	95–99.3%
High-usage	1158.8	20.92	98.2%	95.2–99.3%
High-burst (AT)	1287.8	16.72	98.7%	96.5–99.5%
Compound	2452.3	20.94	99.2%	97.7–99.7%

Asset tracking was found to dominate aggregate demand across all reported regimes, and this dominance became stronger under regimes that stressed AT adoption and burstiness. However, the additional proxy decomposition test indicated that the strength of this dominance was not attributable to the proxy scale alone. A ±30% perturbation of the AT adoption anchor changed Q0.95 by approximately ±29.8%, whereas the stressed AT usage case increased Q0.95 by 74.4%. Therefore, the upper-tail dominance of AT should be interpreted primarily as a usage-uncertainty result, with a material but smaller contribution from proxy-scale uncertainty.

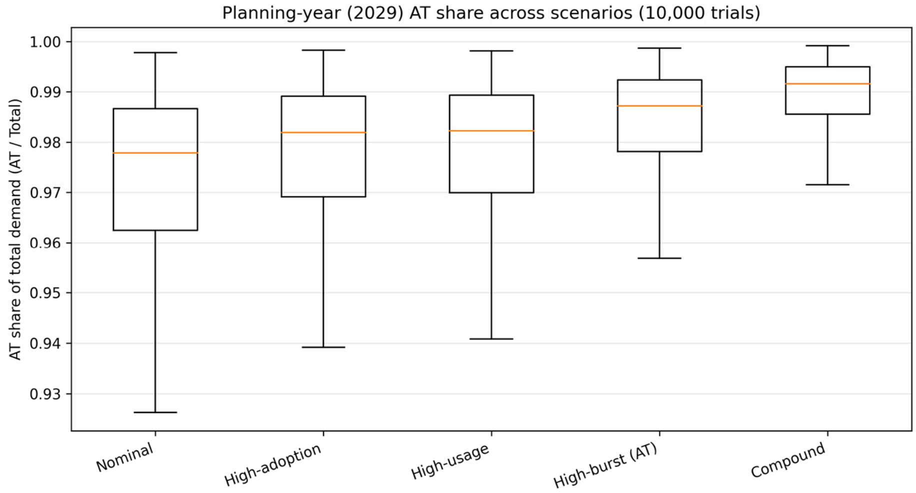


Figure 5. Planning-year (2029) decomposition of total demand into AT and SM components across scenarios.

3.3.3. Sensitivity to the AT Adoption Proxy

Because the AT class was scaled using an M2M-based proxy rather than a direct AT count, an additional decomposition test was performed to separate proxy-scale uncertainty from AT usage uncertainty. The baseline AT adoption path was perturbed multiplicatively by 0.7 and 1.3, while all AT usage priors were kept at nominal settings. A separate stressed-usage case was then evaluated by retaining the baseline AT adoption path and applying the existing stressed AT usage setting. The SM component, Monte Carlo settings, and all other model inputs were left unchanged.

The resulting planning-year summaries are reported in Table 7. Under nominal AT usage, reducing the AT adoption anchor to 0.7× decreased the 2029 median demand from 747.9 to 528.5 Mbit/s and decreased Q0.95 from 2351.8 to 1651.9 Mbit/s, corresponding to a 29.8% reduction in the upper-tail planning metric. Increasing the AT adoption anchor to 1.3× increased the median to 967.2 Mbit/s and increased Q0.95 to 3052.0 Mbit/s, corresponding to a 29.8% increase relative to baseline. By comparison, when the AT adoption anchor was kept at baseline and AT usage priors were stressed, the median increased to 1296.5 Mbit/s and Q0.95 increased to 4101.6 Mbit/s, corresponding to a 74.4% increase relative to baseline.

Table 7. Decomposition of AT proxy-scale and usage-prior effects on planning-year demand (2029).

Scenario	AT Scale Factor	AT Usage Case	Median (Mbit/s)	Q0.95 (Mbit/s)	ΔQ0.95 vs. Baseline
Baseline	1.0	Nominal	747.9	2351.8	0.0%
Low AT anchor	0.7	Nominal	528.5	1651.9	−29.8%
High AT anchor	1.3	Nominal	967.2	3052.0	+29.8%
AT usage stressed	1.0	Stressed	1296.5	4101.6	+74.4%

These results indicate that the reported AT tail dominance was affected materially by the scale choice of the M2M-based proxy, but it was driven more strongly by AT usage uncertainty than by the proxy-scale choice alone. Therefore, the upper-tail result should be interpreted primarily as a usage-driven finding, while a material but smaller contribution from proxy-scale uncertainty should also be acknowledged.

3.3.4. Baseline Alignment for Probabilistic Configurations

The probabilistic summaries for 2029 obtained under the single-model probabilistic baseline (B2) closely match the corresponding summaries from the full configuration used in this application. This alignment indicates that, for the present GB case study settings, scenario multipliers and per-device usage uncertainty account for the dominant share of variability in D_{2029}, while additional stochastic diffusion perturbations (when disabled or weakly expressed) do not materially alter planning-relevant quantiles.

The corresponding B2 scenario-wise quantiles for the 2029 planning year are reported in Table 8.

Table 8. B2 quantiles for D_{2029} by scenario.

Scenario	Median (Mbit/s)	P10–P90 (Mbit/s)	P2.5–P97.5 (Mbit/s)	$Q_{0.95}(D_{2029})$ (Mbit/s)	$\Delta Q_{0.95}$
Nominal	748.3	289–1862.5	179.4–2862	2347	−0.20%
High-adoption	902	346–2250.6	213.1–3460.4	2835.6	−0.38%
High-usage	1178.1	451.6–2940.2	278–4521.4	3704.7	−0.63%
High-burst (AT)	1297.2	492.6–3248.9	300.1–4998.3	4093.2	−0.13%
Compound	2470.8	931.7–6208.9	564.3–9552.8	7824.1	−0.08%

Across scenarios, the B2 and full-method estimates of $Q_{0.95}(D_{2029})$ differ by less than 0.7%, supporting the interpretation that scenario multipliers and per-device usage uncertainty dominate the planning-year tails for this case study configuration.

3.4. Risk-Based Planning Outputs

This subsection translates the predictive distribution of monthly demand in 2029 into decision-relevant risk metrics and a quantile-based capacity recommendation. The primary risk level is set to $\alpha = 0.05$, and capacities are reported in Mbit/s using the conversion in Section 3.3.

Exceedance curves quantify the probability that demand exceeds a candidate capacity C under each scenario:

$$\phi_{2029}(C) = \mathbb{P}(D_{2029} > C) \tag{21}$$

Figure 6 reports $\phi_{2029}(C)$ for the nominal, high-adoption, high-usage, high-burst (AT-focused), and compound regimes. Relative to the nominal regime, adoption and usage stresses shift the curve upward and to the right, indicating higher overload probability for any fixed C. The high-burst regime produces a stronger uplift in the upper tail than the high-usage regime at large C, consistent with burstiness amplifying right-tail risk more than central tendency. The compound regime produces the largest exceedance probabilities across the evaluated capacity range.

A risk-constrained capacity is obtained by selecting the $(1 - \alpha)$ quantile of the planning-year demand distribution:

$$C = Q_{1-\alpha}(D_{2029}), \alpha = 0.05 \tag{22}$$

Table 9 reports C (Mbit/s) for each scenario. Since C is defined by the predictive tail, the relative increase in C across scenarios is larger than the corresponding increase in the median. In particular, the compound regime produces the largest required capacity margin due to the combined effects of stressed adoption and burstiness.

Table 9. Quantile-based capacity $C = Q_{0.95}(D_{2029})$ for $\alpha = 0.05$, by scenario.

Scenario	$\mathbf{Q_{0.95}(D_{2029})}$ (Mbit/s)	C (Tbits/Month)
Nominal	2351.8	6184.8
High-adoption	2846.4	7485.5
High-usage	3728.2	9804.4
High-burst (AT)	4098.8	10,779
Compound	7830.3	20,592.1

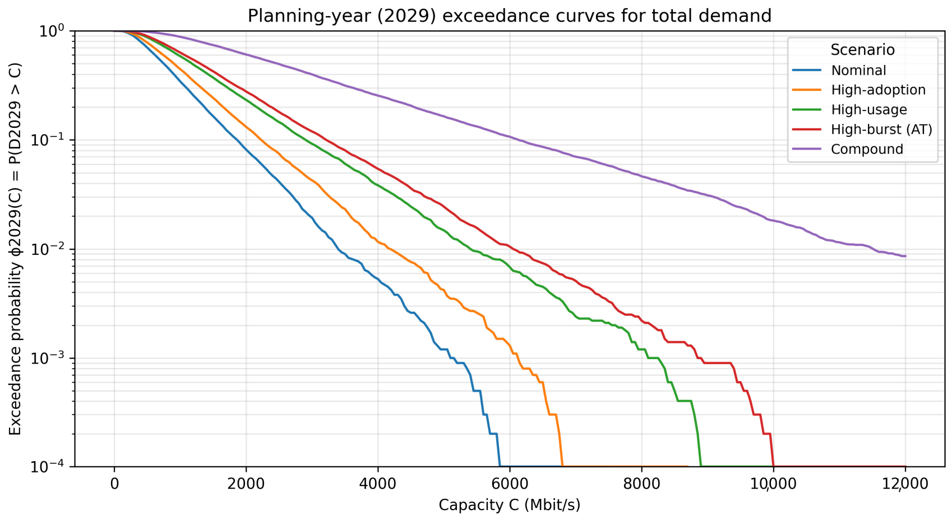


Figure 6. Planning-year (2029) exceedance curves $\phi_{2029}(C)$ for total demand under all scenarios (Mbit/s).

For each scenario, the quantile-based capacity C is the minimum capacity that satisfies $\mathbb{P}(D_{2029} > C) \approx \alpha$ under the forecast distribution. The separation between the median demand and C quantifies the planning margin required to control overload risk. The results indicate that stress regimes affecting AT burstiness produce disproportionately large increases in C, even when median demand changes are moderate, reflecting the sensitivity of quantile-based planning to right-tail behavior.

To support reproducibility, the exceedance curve in Figure 6 is computed empirically from Monte Carlo samples as $\phi_{2029}(C) = K^{-1}\sum_{k=1}^{K} \mathbb{I}\{D_{2029}^{(k)} > C\}$. Additional supporting analyses—busy-hour scaling and expected exceedance $\mathbb{E}[\max(0, D_{2029} - C)]$ for illustrative capacities—are reported separately. The sensitivity of these planning-year risk outputs to the within-period implementation of the adoption path is examined separately in Section 3.5.6.

3.5. Baseline Comparisons and Verification Checks

3.5.1. Deterministic Baseline Trajectory (B1) for 2025–2029

The deterministic baseline (B1) provides a mechanistic reference trajectory $\hat{D}_t$ obtained by fixing diffusion and per-device usage inputs at their nominal point values. For the GB application, B1 is reported for 2025–2029 to summarize the directional evolution of demand under nominal assumptions and to provide a reference against which probabilistic safety margins in 2029 can be interpreted.

B1 values are reported primarily in Mbit/s using the conversion $D_t^{(\mathrm{Mbit/s})} = D_t^{(\mathrm{bits/month})}/(\Delta \cdot 10^6)$ with $\Delta = 2,629,800$ s, with bits/month shown as reference.

The resulting deterministic baseline trajectory for total demand over 2025–2029 is reported in Table 10.

Table 10. B1 deterministic baseline trajectory for total demand (2025–2029).

Year	$\hat{D}_t$ (Mbit/s)	$\hat{D}_t$ (Tbits/Month)
2025	163.3	429.6
2026	186	489.1
2027	209.7	551.5
2028	234.2	615.8
2029	258.9	680.8

3.5.2. Probabilistic Baseline Trajectory (B2) for the Planning Year (2029)

The single-model probabilistic baseline (B2) retains parameter and usage uncertainty while fixing the diffusion family (i.e., diffusion model-form uncertainty is disabled). B2 is summarized for the design year (2029) to provide a probabilistic baseline distribution under nominal assumptions and to contextualize the quantile-based capacity relative to the central tendency.

For planning, the 2029 distribution is the primary focus. The baseline trajectory provides a reference for interpreting how far the quantile-based capacity $C = Q_{0.95}(D_{2029})$ sits above the central tendency relative to earlier years.

The resulting probabilistic baseline distribution for the 2029 planning year is summarized in Table 11.

Table 11. B2 probabilistic baseline (2029).

B2 Scenario	Median (Mbit/s)	P10–P90 (Mbit/s)	P2.5–P97.5 (Mbit/s)	$Q_{0.95}(D_{2029})$ (Mbit/s)
2029	748.3	289–1862.5	179.4–2862	2347

It should be noted that the deterministic baseline B1 fixes inputs at nominal point values (modes), whereas B2 summarizes the full predictive distribution under the same nominal scenario. For the AT class, several priors are intentionally right-skewed (mode close to the lower bound with a long upper tail; e.g., $r_{AT,t}$, $b_{AT,t}$, and s_{AT}, and the multiplicative demand model amplifies this asymmetry. Consequently, the distribution median can exceed the mode-based baseline, and B1 should be interpreted as a "most-likely-value" reference rather than a central estimate.

3.5.3. Empirical Hold-Out Validation and Plausibility

To provide an empirical check against real observed data beyond the scenario-based analysis, an out-of-sample hold-out exercise was conducted for the adoption trajectories used in the GB case study. The purpose of this exercise is to test whether the fitted proxy-adoption paths remain stable when evaluated against later hidden observations and whether the implied growth patterns remain empirically plausible. The validation is, therefore, applied to the observable proxy-adoption fits rather than to the final Monte Carlo demand distribution. For smart metering (SM), the last six quarters of the domestic and non-domestic proxy series were held out, the diffusion models were re-estimated on the earlier training windows, and the resulting forecasts were compared with the hidden observations. For the asset-tracking (AT) class, only a limited two-year hold-out consistency check was conducted on the annual M2M proxy, because the available annual history is short and the series is used only as an AT scale anchor rather than as a direct AT count.

The hold-out results are summarized in Table 12. For SM, very small hold-out errors were obtained. Domestic SM produced a hold-out MAPE of 0.69%, while non-domestic SM produced a hold-out MAPE of 0.70%. The corresponding aggregate SM hold-out MAPE was also 0.69%. In all three SM views, the direction of growth over the hold-out window was reproduced correctly, although the predicted growth was slightly lower than the observed growth. These results indicate that the fitted SM adoption trajectories were stable out of sample and provide strong empirical support for the adoption stage of the model.

Table 12. Hold-out validation results.

Series	Hold-Out Period	MAE	MAPE (%)	Observed Growth Over Hold-Out	Predicted Growth Over Hold-Out	Interpretation
Domestic SM	Last 6 quarters	251,920	0.69	10.09%	8.61%	Strong out-of-sample support
Non-domestic SM	Last 6 quarters	3607	0.70	24.22%	21.69%	Strong out-of-sample support
Aggregate SM	Last 6 quarters	255,528	0.69	10.28%	8.79%	Strong out-of-sample support
AT/M2M proxy	Last 2 years	2,618,503	9.57	2.26%	23.02%	Limited plausibility support

By contrast, the AT hold-out evidence was more limited. The two-year M2M proxy hold-out produced a MAPE of 9.57%, which remained acceptable as a level-error check, but the growth over the hold-out window was overstated: observed growth was 2.26%, whereas predicted growth was 23.02%. Accordingly, the AT result is not interpreted as strong predictive validation. Instead, it is interpreted as a limited plausibility check, consistent with the short annual sample and the manuscript's treatment of the M2M series as an AT scale proxy rather than as a direct AT measurement.

Taken together, the hold-out exercise provides partial empirical validation of the framework through its observable adoption layer, but it does so asymmetrically across classes. Strong out-of-sample support was obtained for SM, whereas the AT evidence should be read more cautiously as plausibility support only. Accordingly, the present results should be interpreted as empirically strengthened at the adoption stage, while the final demand distributions remain scenario-based planning outputs rather than directly validated service traffic forecasts.

3.5.4. Robustness to Correlated AT Usage Shocks

Because AT usage-side assumptions were found to dominate planning-year variability, an additional robustness test was conducted to assess whether the independence assumption materially affected the upper tail of the 2029 demand distribution. The test was intentionally limited in scope and was not treated as a calibrated behavioral dependence model. Instead, three illustrative cases were evaluated for the nominal 2029 total demand forecast: (i) an independent baseline, (ii) a moderate positive dependence case, and (iii) a stronger positive dependence case.

Positive dependence was imposed only among the AT message rate, AT burst factor, and AT activity fraction. The AT payload size was left independent because it was treated as a structural application-level parameter rather than as part of the operational co-movement stress being tested. The tested dependence settings were defined as illustrative robustness inputs rather than as empirically estimated correlation parameters. The target pairwise settings were (0.4, 0.3, 0.2) for the moderate case and (0.6, 0.5, 0.4) for the stronger case, respectively. All adoption paths, usage priors, scenario settings, and trial counts were kept unchanged.

The results are summarized in Table 13. The median was affected only marginally, changing from 767.1 Mbit/s in the independent case to 765.7 Mbit/s under moderate dependence and 760.4 Mbit/s under stronger dependence. By contrast, the upper tail shifted upward materially. Q0.95 increased from 2392.5 Mbit/s in the independent case to 2831.1 Mbit/s under moderate dependence and 2929.9 Mbit/s under stronger dependence, corresponding to increases of 18.3% and 22.5%, respectively. A similar pattern was observed at Q0.99, which increased from 3701.9 Mbit/s to 4476.2 Mbit/s and 4557.7 Mbit/s, corresponding to increases of 20.9% and 23.1%, respectively.

Table 13. Dependence–robustness for the nominal 2029 planning-year demand distribution (Mbit/s).

Case	Corr (Message Rate, Burst Factor)	Corr (Message Rate, Activity Fraction)	Corr (Activity Fraction, Burst Factor)	Median (Mbit/s)	Q0.95 (Mbit/s)	Q0.99 (Mbit/s)	ΔQ0.95 vs. Independent	ΔQ0.99 vs. Independent
Independent baseline	0	0	0	767.1	2392.5	3701.9	0.0%	0.0%
Moderate positive dependence	0.4	0.3	0.2	765.7	2831.1	4476.2	+18.3%	+20.9%
Stronger positive dependence	0.6	0.5	0.4	760.4	2929.9	4557.7	+22.5%	+23.1%

The results indicate that plausible positive co-movement among key AT operational usage inputs does not materially change central tendency, but it does shift the planning-year upper tail upward. Accordingly, the independence assumption should be interpreted as a simplifying baseline rather than as a neutral representation of tail risk. For the present GB application, correlated AT usage shocks provide a more conservative view of planning-year exceedance risk than the independent baseline.

It is noted that minor numerical differences (~1.7%) relative to the main nominal Q0.95 in Table 5 arise from independent random seeds used in the separate dependence–robustness run and do not affect the qualitative conclusions.

3.5.5. Internal Verification and Robustness Checks

Internal verification checks were conducted to support the credibility of the scenario-based application at the level of demand construction, uncertainty propagation, and risk output generation.

- First, dimensional consistency was confirmed: with $r_{c,t}$ specified in messages/device/month and s_c in bits/message, Equation (9) yields D_t in bits/month, and the reported mean-rate equivalents (Mbit/s) follow directly from the stated month-to-seconds conversion $\Delta = 2,629,800$s.
- Second, support constraints and physical bounds were satisfied throughout sampling: the activity factor $a_{c,t}$ was restricted to the interval [0, 1] (with capping after scenario multipliers), the burst factor $b_{c,t}$ remained at or above 1, and demand outputs remained non-negative.
- Third, deterministic recovery was verified: when uncertainty distributions were collapsed to point values, the Monte Carlo implementation reproduced the deterministic baseline $\hat{D}_t$ for the corresponding year within numerical tolerance.
- Fourth, tail behavior under stress was consistent with model intent: under the high-burst and compound regimes, upper quantiles increased more strongly than the median, indicating right-tail amplification driven by burstiness rather than only shifts in central tendency.
- Finally, component accounting was consistent: for each Monte Carlo trial, total demand equaled the sum of AT and SM components (including domestic and non-domestic SM subcomponents where reported), and derived shares (e.g., AT share of total demand) remained within [0, 1]. These checks indicate that the reported planning-year

distributions and derived risk outputs are internally coherent and numerically stable for the chosen sampling size ($K = 10,000$).

The baseline comparisons and verification checks support the interpretation that probabilistic planning outputs for 2029 are driven primarily by scenario regime stresses and per-device usage uncertainty, with quantile-based capacity providing a transparent mapping from forecast tails to a reliability target α. The B1 trajectory provides intermediate-year context without requiring a full probabilistic evaluation for each year.

3.5.6. Interpolation Robustness

Because the smart meter adoption inputs were available at quarterly resolution and the M2M-based asset-tracking scale anchor was available at annual resolution, the baseline implementation treated adoption as piecewise-constant within each reporting period. To test whether this implementation choice materially affected the 2029 planning-year risk outputs, three cases were evaluated. In Case A, the baseline implementation was retained, with smart meter adoption held constant within each quarter and asset-tracking adoption held constant within each year. In Case B, linear monthly interpolation was applied to both smart meter and asset-tracking adoption paths. In Case C, a hybrid diagnostic was evaluated in which only the asset-tracking path was interpolated, while the smart meter path remained piecewise-constant.

In this robustness exercise, the diffusion fits were not re-estimated. The Monte Carlo structure, usage priors, scenario multipliers, and trial count were kept fixed, so that only the deterministic within-period implementation of the adoption path was changed. The comparison was performed on the 2029 planning-year maximum monthly demand distribution, which is more appropriate for this question than the year-end endpoint alone.

Small trial-level differences were observed between the interpolation cases. However, these differences were not sufficient to shift the reported distribution summaries. Across all five scenarios, the median and the 95th percentile of the planning-year maximum monthly demand remained unchanged to reported precision across cases A, B, and C. Accordingly, the percentage differences in the reported summary metrics were classified as negligible.

These results indicate that, for the present GB case study configuration, within-period smoothing of the adoption path may affect some individual simulated outcomes, but it does not materially affect the planning-year risk summaries or the associated risk-based planning conclusions. The piecewise-constant implementation was, therefore, retained in the main analysis, while interpolation was treated as a robustness check rather than as a driver of the headline capacity results.

The quantitative comparison is summarized in Table 14.

Table 14. Interpolation robustness for the 2029 planning-year maximum monthly demand (Mbit/s).

Scenario	Median A	Median B	Median C	Q0.95 A	Q0.95 B	Q0.95 C	Materiality
Nominal	756.3	756.3	756.3	2396.4	2396.4	2396.4	Negligible
High-adoption	911.0	911.0	911.0	2895.8	2895.8	2895.8	Negligible
High-usage	1186.9	1186.9	1186.9	3768.9	3768.9	3768.9	Negligible
High-burst (AT)	1311.4	1311.4	1311.4	4121.0	4121.0	4121.0	Negligible
Compound	2501.1	2501.1	2501.1	7962.3	7962.3	7962.3	Negligible

3.6. Sensitivity Analysis of Planning-Year Results

Sensitivity is evaluated to identify which uncertain assumptions contribute most strongly to the variability of D_{2029}, thereby supporting prioritization of measurement and elicitation effort.

Sensitivity is quantified using the Spearman rank correlation coefficient between each sampled input and the simulated output D_{2029} across the $K = 10,000$ Monte Carlo trials. Rank correlation is appropriate for the present multiplicative demand model (Equation (9)) because it captures monotonic dependence without requiring linearity. The analysis focuses on the uncertain usage inputs $\{r_{c,t}, a_{c,t}, b_{c,t}, s_c\}$; adoption multipliers $(\mu_{m,c}, \mu_{g,c})$ are treated as deterministic scenario stresses rather than sampled quantities.

Table 15 reports Spearman rank correlations for the nominal planning-year distribution. The ordering is stable across scenarios because scenario regimes apply deterministic scaling to the same underlying sampled inputs. The results indicate that planning-year variability is dominated by AT usage characteristics: AT payload size s_{AT} exhibits the strongest association with total demand, followed by the AT burst factor $b_{AT,t}$ and AT message rate $r_{AT,t}$. The AT activity fraction $a_{AT,t}$ shows a smaller, but still positive, contribution. In contrast, SM usage inputs exhibit negligible rank correlation with D_{2029}, consistent with the class decomposition results (Table 6) showing that SM contributes only a small fraction of aggregate demand.

Table 15. Sensitivity of D_{2029} to per-device usage inputs.

Rank	Input	Spearman ρ	Interpretation
1	s_{AT}	0.659	AT payload size is the primary driver of output variability.
2	$b_{AT,t}$	0.519	AT burstiness materially amplifies right-tail uncertainty.
3	$r_{AT,t}$	0.471	AT message rate contributes strongly to demand variability.
4	$a_{AT,t}$	0.125	AT activity fraction has a secondary effect given its narrower support.
5	$b_{SM,t}$	0.012	Negligible influence relative to AT drivers.
6	$a_{SM,t}$	0.010	Negligible influence relative to AT drivers.
7	s_{SM}	−0.007	Negligible influence relative to AT drivers.
8	$r_{SM,t}$	0.002	Negligible influence relative to AT drivers.

These findings reinforce the interpretation that risk-based dimensioning in the GB application is primarily sensitive to AT assumptions, particularly those affecting per-message payload and burstiness. Practically, this suggests that empirical characterization of AT payload distributions and burst dynamics is more consequential for planning-year tail risk than further refinement of SM usage priors under the adopted proxy-based class mapping and scenario structure.

3.7. Case Study Takeaways

The GB application demonstrates that planning-relevant capacity requirements are governed by the right tail of the demand distribution rather than by central tendency. Across the evaluated stress regimes, the median demand increases from nominal to compound, while the upper quantiles increase more strongly under regimes that amplify burstiness and high-usage behavior. The resulting quantile-based capacity recommendations $C = Q_{0.95}(D_{2029})$, therefore, exhibit materially larger scenario separation than median demand, highlighting the importance of reporting predictive distributions and exceedance-based risk metrics for dimensioning. The class decomposition indicates that asset tracking dominates aggregate demand and tail behavior under all regimes, while smart metering contributes a comparatively stable component. The baseline comparisons (B1 and B2) provide intermediate-year context and confirm that, for the present configuration, scenario multipliers and per-device usage uncertainty are the dominant drivers of planning-year tail risk. Sensitivity screening reinforces the decomposition results by showing that planning-year variability is dominated by AT usage assumptions. The strongest drivers of variability are AT payload size, AT burst factor, and AT message rate, while SM usage inputs have negligible influence under the adopted class mapping and priors. In addition, the dependence–robustness test showed that introducing plausible positive co-movement among AT message rate, burstiness, and activity fraction produced little change in the median but materially increased the planning-year upper tail, with Q0.95 increasing by 18.3% under moderate dependence and by 22.5% under stronger dependence. This indicates that independence is a simplifying baseline and that correlated AT operational shocks should be expected to yield a more conservative planning view of exceedance risk.

4. Discussion

This study examined scenario-conditioned forecasting of aggregate IoT traffic demand and the associated capacity risk in a 5G planning context. The discussion concentrates probabilistic interpretation on the end-of-horizon planning year 2029, where full Monte Carlo uncertainty propagation is reported, while intermediate-year trajectories are contextualized through baselines rather than annual predictive distributions. Intermediate-year trajectories are used primarily for context through baselines, while the central planning question is framed in terms of exceedance probability and quantile capacity ($C = Q_{1-\alpha}(D_t)$, $\alpha = 0.05$).

4.1. Main Findings

Across all scenarios, capacity needs are dominated by tail behavior rather than central tendency. For the 2029 planning year, the quantile capacity $C = Q_{0.95}(D_{2029})$ is consistently about 3.1 times the median demand. This stable "tail amplification" shows that planning margins are governed primarily by the right tail of D_{2029}, not by the median (or mean). As a result, reporting median demand alone would materially understate the headroom required to satisfy an exceedance target of $\alpha = 0.05$.

This distinction should be noted clearly. AT payload size emerged as the strongest driver in rank-based sensitivity around the nominal sampling regime, whereas burst-focused scenario stress produced the largest single-axis upward shift in capacity-relevant upper quantiles, and correlated AT usage shocks further increased the tail without materially changing central tendency.

To move beyond a single design percentile, the exceedance curve provides a direct view of the risk–capacity trade-off. Even if planning starts from a reference point, such as $C = Q_{0.95}(D_{2029})$, the curve $\phi_{2029}(C) = P(D_{2029} > C)$ shows how reliability improves as capacity increases. This enables decision-making that is not limited to "meeting $\alpha = 0.05$" but can also evaluate alternative risk postures by observing how quickly exceedance probability declines with incremental additions of C, without the need to re-estimate the demand model.

Scenario comparisons become more decisive when viewed in the tail. Median total demand (Mbit/s) rises from 750.5 (nominal) to 904.6 (high-adoption), 1180.3 (high-usage), 1304.4 (high-burst), and 2473.1 (compound). The same ordering holds at $Q_{0.95}$, but with larger separations; capacity-relevant demand increases from 2351.8 (nominal) to 2846.4 (high-adoption), 3728.2 (high-usage), 4098.8 (high-burst), and 7830.3 (compound). The widening gap between the median and $Q_{0.95}$ underscores that scenario design affects not only typical loading but also the risk profile that ultimately determines dimensioning.

This tail risk is overwhelmingly driven by AT, which accounts for nearly all aggregate demand. The median AT share ranges from 97.8% (nominal) to 99.2% (compound), and the P10–P90 ranges remain high across scenarios (94.0–99.1% in nominal up to 97.7–99.7% in compound). By contrast, SM demand is small and relatively stable, with median totals around 16–21 Mbit/s and only modest increases under high-usage and compound stress. This concentration implies that both aggregate loading and capacity-relevant exceedance risk are primarily governed by AT parameters—especially those controlling usage intensity and burstiness. The sensitivity results provide an independent confirmation of this interpretation. Spearman screening (Section 3.6, Table 15) identifies AT payload size s_{AT} as the strongest driver of variability in D_{2029}, followed by AT burstiness $b_{AT,t}$ and message rate $r_{AT,t}$. This supports prioritizing AT payload/burst characterization when the planning objective is driven by upper-tail risk and quantile-based dimensioning.

Among the single-axis stress tests, burstiness emerges as the more influential lever for upper-tail outcomes. The high-burst (AT) scenario produces a higher $Q_{0.95}(D_{2029})$ (4098.8 Mbit/s) than the high-usage scenario (3728.2 Mbit/s), even though both raise median demand. This indicates that burstiness, captured through $b_{(AT,t)}$, shapes the upper tail more strongly than comparable shifts in typical usage, making it a dominant driver of planning-year risk and a key focus for risk-based capacity planning.

In contrast, additional diffusion or model-form effects play a limited role at the 2029 design point under the current configuration. Across scenarios, $Q_{0.95}(D_{2029})$ differs by less than 0.7% between the B2 baseline and the full configuration. This close alignment suggests that, for the adopted GB configuration and the 2029 endpoint, scenario multipliers and per-device usage uncertainty dominate the capacity-relevant tails, while added diffusion uncertainty/model-form perturbations contribute only marginally to $Q_{0.95}$.

Finally, deterministic baselines help interpret the direction of change over time but do not provide a risk metric for planning. The B1 nominal trajectory increases from 163.3 Mbit/s (2025) to 258.9 Mbit/s (2029), offering an intuitive reference for intermediate-year growth. However, deterministic trajectories do not encode exceedance probabilities or quantile margins, and, therefore, cannot substitute for the probabilistic planning-year outputs used to set capacity under an explicit reliability target.

4.2. Interpretation Relative to the Working Hypotheses

The planning-year results indicate that capacity-relevant outcomes are driven by the right tail of D_{2029} rather than by central tendency. Across all scenarios, $C = Q_{0.95}(D_{2029})$ lies at approximately 3.1× the median, implying that designs anchored on median or mean-rate summaries would systematically understate the headroom required under an exceedance target ($\alpha = 0.05$). The exceedance framing ($\phi_{2029}(C) = \mathbb{P}(D_{2029} > C)$), therefore, provides decision-relevant information that is not recoverable from point forecasts alone.

Furthermore, the Monte Carlo decomposition shows that AT contributes nearly all aggregate demand across regimes, with the median AT share ranging from 97.8% (nominal) to 99.2% (compound), and high shares are maintained across trials (P10–P90 ranges from 94.0–99.1% in nominal to 97.7–99.7% in compound). This concentration implies that the planning-year exceedance probability and quantile capacity are structurally governed by AT-related terms in $D_t = \sum_c N_{c,t}\, a_{c,t} r_{c,t} s_c b_{c,t}$, whereas the SM contribution remains comparatively small and stable (median totals around 16–21 Mbit/s). Under the studied configuration, refinements to SM assumptions are, therefore, expected to have limited impact on aggregate tail risk unless accompanied by substantial changes in SM scale or intensity.

Among single-axis stresses, the high-burst (AT) regime yields a higher $Q_{0.95}(D_{2029})$ than the high-usage regime, despite both producing elevated medians. This ordering is consistent with the multiplicative role of burstiness through $b_{AT,t}$, which amplifies high-demand realizations more strongly than comparable shifts in typical usage. The compound regime, combining multiple stresses, produces the largest separation at $Q_{0.95}$, reinforcing that tail risk is particularly sensitive to the joint interaction of adoption, usage intensity, and burstiness.

Moreover, the scenario set generates coherent, interpretable separation across demand distributions and the associated quantile capacities, providing a transparent mapping from plausible futures to exceedance-based risk outputs. Additional credibility was provided in the analysis by both the adoption-stage hold-out checks and the dependence–robustness test. Strong out-of-sample support was obtained for the SM adoption trajectories, whereas the AT proxy yielded only limited plausibility support. In addition, when plausible positive dependence was introduced among AT message rate, burstiness, and activity fraction, the 2029 upper tail shifted upward materially while the median changed only marginally. This indicates that independence is useful as a baseline simplification, but it is somewhat optimistic for tail-risk interpretation. In this framing, scenarios and robustness variants function as policy-relevant stress levers that translate directly into capacity targets under a specified reliability threshold.

Last, across scenarios, $Q_{0.95}(D_{2029})$ differs by less than 0.7% between the B2 baseline and the full configuration, indicating that scenario multipliers and per-device usage uncertainty are the principal contributors to the planning-year tail. This finding is endpoint- and configuration-specific. It implies that, for 2029 dimensioning under the adopted priors and perturbations, additional diffusion uncertainty/model-form effects contribute marginally to $Q_{0.95}$ relative to the imposed scenario stresses and usage variability.

Taken together, the hypothesis assessments support a risk-based planning logic in which exceedance targets define the relevant operating point, AT assumptions dominate aggregate outcomes, and burstiness constitutes the most influential tail-shaping mechanism among the examined single-axis stressors.

4.3. Comparison with Prior Studies

Prior work relevant to the present results spans three intersecting strands: diffusion-based forecasting of technology adoption, uncertainty quantification and Monte Carlo propagation for forecasting and engineering design, and traffic engineering for IoT/vertical services, especially the role of burstiness and tail risk in capacity planning. The 2029 outcomes connect naturally to these strands, while also clarifying how the adopted scope and reporting choices differ from common approaches in the literature.

S-curve diffusion models (and related variants) are widely used to represent adoption trajectories when saturation effects and growth dynamics matter, and when explanatory covariates are limited, unstable, or unavailable. The present results align with this tradition insofar as adoption-oriented stress (high-adoption) lifts aggregate demand relative to nominal, and—more importantly—interacts with usage and burstiness stresses to create stronger separation in the planning-relevant tail. At the same time, the analysis is anchored on an endpoint design decision rather than year-by-year distributional tracking. In this endpoint setting, the near-equivalence between B2 and the full configuration at $Q_{0.95}(D_{2029})$ indicates that scenario multipliers and usage uncertainty dominate tail outcomes at the horizon for the adopted configuration. This is compatible with diffusion studies that emphasize identifiability challenges and model-form uncertainty earlier in a technology lifecycle while suggesting a weaker influence on endpoint demand once saturation anchors and scenario conditioning constrain the plausible adoption range [7–9,16].

A second body of related work comes from probabilistic forecasting in engineering and operations, where Monte Carlo simulation is increasingly used to propagate parameter and model uncertainty into decision-relevant quantities. The stable ratio between $Q_{0.95}(D_{2029})$ and the median across scenarios reflects a typical property of multiplicative demand constructions; compounding uncertainty in scale, activity, and burstiness yields right-skewed distributions in which tail quantiles substantially exceed central summaries. Relative to studies that report prediction intervals around point forecasts, the exceedance and quantile capacity framing used here is closer to reliability-based design. In particular, α directly expresses a tolerated shortfall probability, and $C = Q_{1-\alpha}(D_{2029})$ converts probabilistic demand into an explicit capacity requirement [11,12,14,15].

The results also match established findings in IoT traffic engineering, where heterogeneity across device types and applications is often pronounced and aggregate load can be shaped by a small number of high-activity or highly synchronous sources. In such settings, burstiness and correlation effects can generate heavy (or long) right tails, so peak behavior becomes disproportionately important for planning. The scenario outcomes reflect this mechanism clearly; among single-axis stresses, the high-burst regime produces a larger $Q_{0.95}(D_{2029})$ than high-usage, indicating that burstiness (through $b_{(AT,t)}$) is an especially influential driver of tail risk. This aligns with broader traffic engineering evidence that peak-to-average behavior, synchronization, and protocol dynamics can dominate capacity margins, even when average rates remain moderate. The dominance of AT, both in median contribution and in trial-to-trial share, further mirrors the empirical observation that a limited subset of service types can govern aggregate demand and its uncertainty [5,6,10].

Within communications and infrastructure planning, percentile-based targets, service-level constraints, and exceedance metrics are commonly used to manage demand variability and uncertainty. The use of $C = Q_{0.95}(D_{2029})$ fits squarely within this tradition, with $\alpha = 0.05$ representing a tunable policy choice that encodes the desired risk posture. Compared with deterministic busy-hour multipliers or single-value headroom factors, quantile-based dimensioning provides an explicit, auditable link between the assumed uncertainties and the resulting capacity requirement. Moreover, exceedance curves $\phi_{2029}(C)$ give a more informative view than any single percentile by showing how rapidly exceedance probability declines as capacity is incrementally increased.

Two departures from much of the prior literature are particularly important for interpreting the contribution. First, probabilistic reporting is concentrated on the design year (2029), while intermediate years are presented primarily for context through baselines. This differs from studies that provide full predictive distributions across all horizons, but it matches planning processes that treat the end-of-horizon year as the principal dimensioning milestone. Second, signaling load S_t is not instantiated in the GB application. While signaling is treated explicitly in parts of the IoT performance literature, the present scope intentionally leaves S_t to practitioners in order to keep the main analysis focused on demand uncertainty and capacity risk.

Positioning Relative to Scenario-Based Optimization Frameworks

The present framework differs fundamentally from scenario-based mathematical optimization approaches, such as mixed-integer linear and nonlinear programming models. For example, in [31], a long-term stochastic scenario-based optimization model is formulated with explicit decision variables, feasibility constraints, and an optimization objective over a formally specified infrastructure design problem. Such frameworks can provide solver-based optimality for the chosen objective and constraints, but they usually require a much more fully specified design space, scenario set, and cost structure than are available in early-stage or data-lean planning settings.

By contrast, the present study does not solve an infrastructure optimization problem. Its purpose is to generate probabilistic demand distributions, exceedance curves, and quantile-based planning rules under uncertain adoption and usage. In design terms, the framework trades formal optimality for transparency, lighter data requirements, and easier reproducibility because the assumptions enter through auditable diffusion fits, bounded usage priors, scenario multipliers, and Monte Carlo sampling rather than through a large solver-based model. With respect to scalability, the present workflow is also easier to deploy when the immediate planning need is to quantify demand risk before a detailed optimization model becomes feasible. The two approaches should, therefore, be seen as complementary; the present framework provides a forecasting and risk-dimensioning layer that can inform, but does not replace, a downstream MILP/MINLP network optimization stage.

4.4. Practical Implications for Operators and Regulators

The planning-year results translate into operationally meaningful implications because the adopted risk outputs—exceedance probability and quantile capacity—map directly to dimensioning decisions and governance choices. Taken together, the findings show that aggregate IoT demand planning is fundamentally a tail management problem, and that the highest-impact levers are those that shape AT intensity and burst behavior.

A key implication is that risk-based capacity targets make planning decisions explicit and auditable. Defining capacity as $C = Q_{1-\alpha}(D_{2029})$ operationalizes a service objective in terms of an exceedance tolerance α, providing a transparent basis for internal governance (for example, engineering approval thresholds) and external accountability (including regulatory scrutiny of coverage and performance commitments). Exceedance curves $\phi_{2029}(C) = P(D_{2029} > C)$ extend this transparency by supporting sensitivity analysis around any proposed capacity level, enabling concrete assessments of how incremental capacity reduces risk rather than relying on implicit headroom factors.

Because tail outcomes dominate, conservative dimensioning cannot be inferred from median demand. The persistent gap between medians and $Q_{0.95}$ across scenarios implies that planning based on central tendency will materially under-provision when requirements are expressed in exceedance terms. For operators, this motivates embedding percentile-based targets in annual planning cycles and aligning investment gates with an explicitly chosen α. For regulators, it suggests that evaluation criteria focused only on average traffic or typical-day performance can overlook the conditions under which service objectives fail—precisely the conditions that drive capacity sufficiency.

The results also indicate that AT-focused interventions offer the highest leverage for reducing aggregate risk. Since AT contributes the overwhelming share of demand across trials, mitigation options that change AT activity and burstiness should dominate the planning toolkit. Practical examples include traffic shaping and policy control that reduce effective burstiness (as reflected through $b_{(AT,t)}$), admission control and prioritization for latency-critical or high-volume AT applications, and service-level agreements with verticals that constrain peak behavior, synchronization, or reporting cadence. By comparison, marginal refinements to SM assumptions are unlikely to shift aggregate tail risk meaningfully under the studied configuration.

Scenario stress testing further supports robust planning when uncertainty is partly structural rather than purely statistical. The mismatch between available proxy series and the target AT class implies that uncertainty is also epistemic—stemming from imperfect observability and model structure. Scenario-conditioned planning addresses this by exploring plausible adoption and intensity regimes rather than relying on a single calibrated forecast. Within this logic, the compound scenario functions as an upper-stress envelope for contingency planning, while the single-axis scenarios help attribute which levers—adoption, usage, or burstiness—are responsible for increases in exceedance risk.

These conclusions also inform portfolio planning and the timing of upgrades. Concentrating full uncertainty propagation on the design year matches planning processes that treat end-of-horizon requirements as the primary trigger for investment decisions. Baseline trajectories for intermediate years remain useful for budgeting and scheduling (for example, managing upgrade lead times and procurement staging), while the 2029 quantile capacity provides the sizing criterion that prevents deferring investment until tail risk becomes operationally visible.

https://doi.org/10.3390/forecast8020028

At the policy level, a quantile-based framing supports clearer national coverage objectives and more consistent capacity oversight. Separating the reliability target α from the traffic forecast enables more coherent comparisons across regions and operators, since differences in outcomes can be interpreted in light of an explicit risk posture. It also clarifies that capacity sufficiency hinges on tail conditions shaped by vertical-specific behavior—such as synchronized updates or event-driven reporting—which may motivate sector coordination, best-practice guidance, or incentive schemes aimed at reducing extreme burst patterns.

For practical reporting, expressing demand and capacity in mean-rate units (Mbit/s) integrates cleanly with engineering workflows and avoids ambiguity around calendar-month length, but traceability to the underlying bits/month demand should be preserved for auditability and cross-study comparison. The adopted average-month conversion ($\Delta = 2,629,800$ s) provides a consistent translation between reporting forms without changing the underlying demand definition.

Finally, the framework is transferable across regions and operators with limited changes, provided local inputs are substituted in the demand identity. Reuse primarily requires an appropriate proxy or scenario pathway for $N_{(c,t)}$ (via the class mapping $N_{(c,t)} = A_{(c,t)}$) and locally defensible assumptions or priors for $a_{(c,t)}$, $r_{(c,t)}$, and $b_{(c,t)}$ (with s_c specified consistently). Because outputs are risk-indexed, the reliability target α should always be stated explicitly as a policy choice, enabling direct comparison of $C_t = Q_{1-\alpha}(D_t)$ across operators, even when proxies and calibration choices differ.

Illustrative Mapping of Aggregate Mean-Rate Output to Engineering Planning

The reported Mbit/s values should be interpreted as aggregate planning outputs rather than as direct cell-level capacity prescriptions. A minimal engineering mapping, therefore, requires an additional, deployment-specific translation step. For illustration, if the nominal 2029 quantile-based capacity is taken as $Q_{0.95}(D_{2029}) = 2351.8$ Mbit/s and a busy-hour scaling factor of $\kappa_{BH} = 2$ is assumed, the corresponding busy-hour offered load is approximately 4.70 Gbit/s. If that load is then distributed across 100 effective serving cells, the implied average per-cell busy-hour offered load is approximately 47 Mbit/s before any radio, scheduler, or control-plane modeling is applied.

This example is illustrative only. The actual mapping depends on deployment geometry, site density, sectorization, scheduler behavior, traffic synchronization, signaling/control-plane constraints, and the operator's engineering rules. Accordingly, the present framework should be interpreted as providing an aggregate demand risk input to engineering dimensioning rather than a complete cell-level design procedure.

4.5. Limitations and Threats to Validity

Several limitations bound how the reported demand distributions and quantile-based capacities were interpreted and generalized. These limitations arise mainly from the way adoption proxies are constructed, the level of aggregation at which demand is modeled, constraints on external validation, and deliberate scoping choices made to keep probabilistic reporting tractable and decision-facing.

A primary threat to validity is proxy mismatch and class mapping risk, especially for the AT class. Although the model maps class population to adoption via $N_{c,t} = A_{c,t}$ and evaluates demand through $D_t = \sum_c N_{c,t}\, a_{c,t}\, r_{c,t}\, s_c\, b_{c,t}$, the available proxy used to calibrate AT may not reflect the full diversity of 5G AT applications. In particular, it may miss heterogeneity in concurrency patterns, traffic shaping, and peak-to-average characteristics. Because the upper tail of D_t is strongly influenced by burstiness $b_{AT,t}$ and usage variability, any proxy–class mismatch can bias not only expected levels but also tail behavior, which in turn affects exceedance probabilities and quantile capacity estimates.

A second limitation concerns external validation of forecasting accuracy at the final demand level. Independent out-of-sample 5G service traffic measurements were not available, and, therefore, predictive accuracy for aggregate demand is not claimed in the conventional end-to-end statistical sense. At the same time, the GB application does include a meaningful empirical hold-out check based on actual observed proxy data for the adoption stage of the framework. Strong out-of-sample support was obtained for the smart meter proxy trajectories, whereas only a limited plausibility check was possible for the AT/M2M proxy because the annual series is short and serves only as an AT scale anchor rather than as a direct AT measurement. Consequently, the empirical evidence should be interpreted as support for the stability and realism of the observable adoption layer rather than as full validation of the final demand distribution. This still materially strengthens the case study relative to a purely scenario-based exercise, but continued caution remains necessary when interpreting the AT adoption pathway and the tail-risk results that depend on it.

Relatedly, the framework operates at an aggregate planning level. Demand is reported as a mean-rate equivalent in Mbit/s (with bits/month retained as a traceable reference), which is appropriate for high-level capacity planning and risk framing. However, this abstraction does not capture cell-level scheduling dynamics, spatial heterogeneity, or interference-limited behavior in the radio access network. Consequently, the quantile capacity $C = Q_{0.95}(D_{2029})$ should be interpreted as an aggregate planning requirement rather than a direct prescription for site- or cell-level engineering.

The probabilistic reporting is also shaped by an endpoint planning-year focus. Full Monte Carlo uncertainty propagation is presented for the planning year (2029), while intermediate years are contextualized through baselines. This design matches common endpoint-driven planning, but it limits insight into how uncertainty evolves across the horizon and whether earlier-year tails exhibit different sensitivities or driver rankings. Accordingly, the observation that the baseline configuration and the full configuration are near-equivalent at $Q_{0.95}(D_{2029})$ is specific to the endpoint emphasis and the adopted perturbation structure; it should not be assumed to hold uniformly for intermediate years or alternative uncertainty specifications.

Beyond reporting choices, the model includes deliberate simplifications and omitted mechanisms. The demand construction is multiplicative and intentionally parsimonious, which improves interpretability and auditability but omits many operational mechanisms that can matter in practice, such as episodic event effects, protocol-driven synchronization, and richer forms of dependence across traffic drivers. This issue was examined through a limited robustness test in which positive dependence was introduced among the AT message rate, burst factor, and activity fraction. That test showed a material upward shift in Q0.95 and Q0.99, even though the median changed little. Accordingly, independence should be interpreted as a simplifying baseline rather than as a fully neutral assumption for tail-risk planning. Moreover, the signaling proxy is defined conceptually but not instantiated in the GB application and is left to practitioners. This limits what can be concluded about control-plane load and its coupling to data-plane burstiness in the present case study. Additionally, busy-hour conversion is represented with a scalar uplift applied to the monthly mean-rate equivalent. This is a simplification; many IoT applications exhibit clock-driven synchronization (e.g., scheduled reporting windows) that can create short, intense bursts that are not well captured by a single scalar. The scenario matrix partially mitigates this via burst/duty stress (especially for AT), but deployment studies should validate busy-hour behavior against service-specific telemetry when available.

There are also reporting conventions and conversion assumptions that should be acknowledged. The conversion from bits/month to Mbit/s relies on an average-month convention ($\Delta = 2,629,800$ s). While this improves comparability across months, real calendar variation can introduce small differences in mean-rate interpretation when comparing reported values to measurements computed over specific months. This is unlikely to change scenario ordering or the qualitative finding that tail risk is dominated by AT usage and burst behavior, but it may matter for precise benchmarking. Also, some supporting analyses are deferred. Busy-hour scaling and the expected exceedance metric $\mathbb{E}[\max(0, D_t - C)]$ are left for supplementary work. As a result, the main text emphasizes exceedance probability and quantile capacity as the primary decision outputs rather than operationally specific peak-hour engineering margins or cost-of-shortfall interpretations.

4.6. Future Research

Several extensions could increase the practical fidelity of the framework while preserving the interpretability of the demand identity $D_t = \sum_c N_{(c,t)}\ a_{(c,t)}\ r_{(c,t)}\ s_c\ b_{(c,t)}$ and the core risk outputs $\phi_t(C) = P(D_t > C)$ and $C_t = Q_{1-\alpha}(D_t)$. Broadly, these directions focus on improving measurement and proxies, strengthening the operational mapping from aggregate demand to engineering dimensioning, and enriching how risk is characterized and communicated.

A first priority is improved AT representation and proxy development. Given AT's dominant contribution to both the level of demand and the upper-tail risk, targeted work on better AT proxies and calibration sources would likely yield the largest gains. Future studies could combine multiple indicators—such as vertical subscription records, device inventories, enterprise contracts, slice-level counters, and application-layer telemetry where available—to reduce structural mismatch between proxies and the target AT class. In parallel, explicitly modeling heterogeneity within AT by segmenting it into sub-classes with distinct $a_{(c,t)}$, $r_{(c,t)}$ and $b_{(c,t)}$ would improve tail realism without changing the overall demand identity.

A second extension is explicit signaling load modeling (S_t). While S_t is left to practitioners in the current application, incorporating a signaling proxy into the same probabilistic pipeline would allow joint assessment of user-plane and control-plane stress. This is particularly relevant in IoT deployments where attach, paging, and periodic reporting behaviors can generate disproportionate control-plane load relative to user-plane throughput. A practical research direction is to define S_t in a way that aligns with available counters and then quantify how signaling depends on burstiness and synchronization mechanisms.

Busy-hour scaling can also be brought into the main risk layer to better align outputs with standard engineering KPIs. In the current framing, busy-hour factors are treated as supporting analysis; integrating them directly would enable busy-hour exceedance curves and busy-hour quantile capacities. One approach is to model the busy-hour factor as an uncertain multiplier—potentially class- and scenario-dependent—and propagate it jointly with D_t, maintaining traceability back to the monthly definition while producing results in the terms most commonly used for dimensioning.

Beyond probability-of-exceedance and quantile capacity, future reporting could include expected exceedance and cost-of-shortfall interpretations. Extending the outputs to $E[\max(0, D_t - C)]$ would provide an operationally meaningful measure of how much demand is expected to exceed a proposed capacity level, which supports economic interpretation (for example, expected unmet demand) and cost-aware optimization of α and investment timing. This is especially helpful when stakeholders need to trade capital expenditure against quantified shortfall risk rather than relying on a single reliability threshold.

Temporal structure and seasonality are another area for extension when justified by evidence. The current scope uses a monthly resolution with an average-month conversion and does not add seasonal dynamics. Where local measurements show strong seasonality or event-driven bursts, future work could incorporate seasonal structure into $N_{(c,t)}$ or into the intensity terms $\left(a_{(c,t)}, r_{(c,t)}, b_{(c,t)}\right)$ while preserving monthly accounting for D_t. The key is to avoid over-parameterization when historical data are limited, so any additional structure should be supported by clear empirical signals.

Relatedly, correlation and synchronization effects across devices could be made explicit. The multiplicative construction captures variability through distributions on the component terms, but it does not directly represent cross-device dependence or synchronized behavior such as firmware updates, alarms, or market-wide events. Introducing correlated shocks—via copula-based dependence structures or scenario-defined common factors—would likely improve tail realism and strengthen the link between operational events and exceedance risk.

An important practical step is mapping aggregate demand to RAN and transport dimensioning. The aggregate output C_t is not, by itself, a cell-level prescription. Future work can connect D_t and $\phi_t(C)$ to spatial models that allocate demand across sites, incorporate coverage and interference constraints, and translate mean-rate equivalents into schedulable resources under realistic traffic assumptions. This would allow uncertainty to propagate consistently from adoption and usage assumptions to engineering outcomes such as site upgrades, spectrum requirements, and backhaul provisioning.

Finally, governance can be strengthened through multi-α reporting. Although $\alpha = 0.05$ is the primary design choice here, reporting a small set of α values (for example, policy-relevant percentiles) would better support processes that distinguish baseline service objectives from contingency planning. The exceedance curve view naturally enables this: decision-makers can read risk at multiple capacity points without changing the underlying simulation or re-estimating the model.

5. Conclusions

A probabilistic diffusion-to-dimensioning workflow was proposed for medium-term 5G/NB-IoT planning under simultaneous uncertainty in adoption and per-device behavior. The framework was designed to convert uncertain diffusion and usage assumptions into predictive demand distributions, exceedance curves, and quantile-based capacity rules, rather than into deterministic point forecasts or optimization-based infrastructure decisions.

In the GB application, the 2029 quantile-based capacity at $\alpha = 0.05$ ranged from 2351.8 Mbit/s under the nominal regime to 7830.3 Mbit/s under the compound regime, while the required quantile capacity remained approximately 3.1 times the median demand across scenarios. This indicates that design-year capacity is governed by upper-tail conditions rather than by central tendency. The supporting analyses further clarified the source of this tail risk. A $\pm 30\%$ perturbation of the AT adoption anchor changed $Q_{0.95}$ by approximately $\pm 29.8\%$, whereas stressed AT usage increased $Q_{0.95}$ by 74.4%. Plausible positive dependence among key AT operational inputs increased $Q_{0.95}$ by 18.3–22.5%, while interpolation of the quarterly and annual adoption paths did not materially change the planning-year summaries. In addition, the difference between the B2 baseline and the full configuration remained below 0.7% at $Q_{0.95}$, indicating that planning-year tails were driven primarily by scenario multipliers and per-device usage uncertainty rather than by additional diffusion perturbations under the adopted configuration.

The evidence also strengthened the empirical grounding of the framework, although asymmetrically across classes and stages. Strong out-of-sample support was obtained for the smart meter adoption stage using actual hold-out observations, whereas the AT/M2M proxy provided plausibility support only, which is consistent with the shorter annual history and the broader scope of the proxy relative to the AT class. Accordingly, the upper-tail dominance of AT should be interpreted primarily as a usage-driven finding, with a material but smaller contribution from proxy-scale uncertainty, and with empirical support that is stronger for the observable adoption layer than for the final aggregate demand output.

Two practical implications follow. First, medium-term 5G IoT planning should be based on exceedance curves and percentile-based capacity targets rather than on deterministic point sizing because average or median demand materially understates required headroom. Second, uncertainty reduction efforts should be focused on AT usage characterization, especially payload, burst behavior, and message intensity, since these inputs exert the strongest influence on capacity-relevant outcomes.

The framework should also be interpreted within its scope. It provides an aggregate forecasting and risk-dimensioning tool for data-lean planning settings; it is not a cell-level engineering model, and it is not a scenario-based MILP/MINLP optimization framework. Future work should, therefore, extend the present demand-risk layer toward richer empirical telemetry, spatial allocation, signaling-aware dimensioning, and downstream optimization models that translate aggregate risk outputs into network design decisions.

Author Contributions: Conceptualization, D.K. and D.V.; methodology, N.K.; validation, N.K. and D.K.; formal analysis, N.K.; investigation, N.K.; writing—original draft preparation, N.K.; writing—review and editing, D.K. and D.V.; visualization, N.K.; supervision, D.K. and D.V.; project administration, D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. Smart metering data are available from the Department for Energy Security and Net Zero (DESNZ) at https://www.gov.uk/government/collections/smart-meters-statistics (accessed on 13 January 2026); M2M mobile network subscription data are available from Ofcom (2012–2015) at https://www.ofcom.org.uk/research-and-data/telecoms-research/data-updates (accessed on 13 January 2026) and the International Telecommunication Union (ITU) (2016 onward) at https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx (accessed on 13 January 2026).

Acknowledgments: During the preparation of this work, the authors used ChatGPT 5.4 in order to improve the clarity and fluency of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3GPP	Third Generation Partnership Project
5G	Fifth Generation (Mobile Networks)
AIC	Akaike Information Criterion
AT	Asset Tracking
CIoT	Cellular Internet of Things
CVaR	Conditional Value-at-Risk
DESNZ	Department of Energy Security and Net Zero
GB	Great Britain
IoT	Internet of Things
ITU	International Telecommunication Union
KPI	Key Performance Indicator
M2M	Machine to Machine

mMTC	Massive Machine-Type Communications
NB-IoT	NarrowBand Internet of Things
RAN	Radio Access Network
SM	Smart Metering

References

1. ITU-R. *IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*; Recommendation ITU-R M.2083-0; ITU: Geneva, Switzerland, 2015; Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/r-rec-m.2083-0-201509-i!!pdf-e.pdf (accessed on 13 January 2026).
2. 3GPP. *Study on Scenarios and Requirements for Next Generation Access Technologies*; Technical Report (TR) 38.913, V16.0.0; 3GPP: Sophia Antipolis, France, 2020. Available online: https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/16.00.00_60/tr_138913v160000p.pdf (accessed on 13 January 2026).
3. ITU-T. *Overview of the Internet of Things; Recommendation ITU-T Y.2060 (06/2012)*; ITU: Geneva, Switzerland, 2012. Available online: https://www.itu.int/rec/dologin_pub.asp?id=T-REC-Y.2060-201206-I!!PDF-E&lang=e&type=items (accessed on 13 January 2026).
4. 3GPP. *Study on RAN Improvements for Machine-Type Communications*; Technical Report (TR) 37.868, V11.0.0; 3GPP: Sophia Antipolis, France, 2011.
5. Shafiq, M.Z.; Ji, L.; Liu, A.X.; Pang, J.; Wang, J. A First Look at Cellular Machine-to-Machine Traffic: Large Scale Measurement and Characterization. In Proceedings of the ACM SIGMETRICS 2012, London, UK, 11–15 June 2012. [CrossRef]
6. Laner, M.; Svoboda, P.; Nikaein, N.; Rupp, M. Traffic Models for Machine Type Communications. In Proceedings of the 10th International Symposium on Wireless Communication Systems (ISWCS 2013), Ilmenau, Germany, 27–30 August 2013.
7. Bass, F.M. A New Product Growth for Model Consumer Durables. *Manag. Sci.* **1969**, *15*, 215–227. [CrossRef]
8. Meade, N.; Islam, T. Modelling and Forecasting the Diffusion of Innovation—A 25-Year Review. *Int. J. Forecast.* **2006**, *22*, 519–545. [CrossRef]
9. Mahajan, V.; Muller, E.; Bass, F.M. New Product Diffusion Models in Marketing: A Review and Directions for Research. *J. Mark.* **1990**, *54*, 1–26. [CrossRef]
10. ITU-T. *Terms and Definitions of Traffic Engineering; Recommendation ITU-T E.600 (03/1993)*; ITU: Geneva, Switzerland, 1993. Available online: https://www.itu.int/rec/dologin_pub.asp?id=T-REC-E.600-199303-I!!PDF-E&lang=e&type=items (accessed on 13 January 2026).
11. Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341. [CrossRef] [PubMed]
12. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*, 2nd ed.; Springer: New York, NY, USA, 2004; ISBN 978-0387212395.
13. Rockafellar, R.T.; Uryasev, S. Conditional Value-at-Risk for General Loss Distributions. *J. Bank. Financ.* **2002**, *26*, 1443–1471. [CrossRef]
14. Morgan, M.G.; Henrion, M. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*; Cambridge University Press: Cambridge, UK, 1990.
15. Vose, D. *Risk Analysis: A Quantitative Guide*, 3rd ed.; John Wiley & Sons: Chichester, UK, 2008.
16. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
17. Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications*, 6th ed.; Springer: Berlin/Heidelberg, Germany, 2003.
18. Kanellos, N.; Katsianis, D.; Varoutas, D. On the Introduction of Diffusion Uncertainty in Telecommunications' Market Forecasting. *Eng. Proc.* **2021**, *5*, 13. [CrossRef]
19. Shapiro, S.S.; Wilk, M.B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52*, 591–611. [CrossRef]
20. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef]
21. Sobol', I.M. Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates. *Math. Comput. Simul.* **2001**, *55*, 271–280. [CrossRef]
22. Saltelli, A. Making Best Use of Model Evaluations to Compute Sensitivity Indices. *Comput. Phys. Commun.* **2002**, *145*, 280–297. [CrossRef]
23. Duan, J.-H.; Li, W.; Zhang, X.; Lu, S. Forecasting Fine-Grained City-Scale Cellular Traffic with Sparse Crowdsourced Measurements. *Comput. Netw.* **2022**, *214*, 109156. [CrossRef]
24. Li, S.; Magli, E.; Francini, G.; Ghinamo, G. Deep Learning Based Prediction of Traffic Peaks in Mobile Networks. *Comput. Netw.* **2024**, *240*, 110167. [CrossRef]
25. Yang, Y.; Geng, S.; Zhang, B.; Wang, Z.; Zhang, Y.; Doermann, D.; Zhang, J. Long Term 5G Network Traffic Forecasting via Modeling Non-Stationarity with Deep Learning. *Commun. Eng.* **2023**, *2*, 33. [CrossRef]
26. García-Martín, J.P.; Torralba, A. Energy Consumption Analytical Modeling of NB-IoT Devices for Diverse IoT Applications. *Comput. Netw.* **2023**, *232*, 109855. [CrossRef]

27. Department for Energy Security and Net Zero (DESNZ). Smart Meters in Great Britain, Quarterly Update (Official Statistics). Available online: https://www.gov.uk/government/statistics/smart-meters-in-great-britain-quarterly-update-september-2025 (accessed on 13 January 2026).
28. Ofcom. The Communications Market 2016—UK Telecoms. Available online: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/cmr/cmr16/uk/uk_telecoms.pdf (accessed on 13 January 2026).
29. ITU. ITU World Telecommunication/ICT Indicators with Definitions. 14 August 2017. Available online: https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/Documents/Events/2017/Aug-aspstat2017/WTI_Indicators_with_Definitions_14Aug2017_final.pdf (accessed on 13 January 2026).
30. ITU. List of Indicators Included in the World Telecommunication/ICT Indicators Database (WTID). Available online: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/wtid/WTID2017_ListOfIndicators_Dec.pdf (accessed on 13 January 2026).
31. Fathollahi, A.; Derakhshandeh, S.Y.; Ghiasian, A.; Masoum, M.A.S. Optimal Siting and Sizing of Wireless EV Charging Infrastructures Considering Traffic Network and Power Distribution System. *IEEE Access* **2022**, *10*, 117105–117117. [CrossRef]