Addressing Uncertainty in Multi-Step Traffic-Driven Service Provisioning

Hafsa Maryam, Giannis Savva, Georgios Ellinas, and Tania Panayiotou

***Abstract*—Network traffic prediction enables proactive service provisioning in telecommunication networks. However, prediction accuracy alone does not capture the impact of errors on provisioning decisions and potential quality-of-service (QoS) violations. To address this, both model and input uncertainty must be explicitly quantified and incorporated into decision-making. This work investigates uncertainty estimation for multi-step-ahead traffic predictions and proposes an uncertainty-aware, multi-period service provisioning framework that improves spectrum utilization while mitigating service disruptions. Uncertainty is estimated using Monte Carlo dropout and compared with a baseline margin-based approach, highlighting the importance of accurate uncertainty quantification, especially in terms of spectrum savings.**

I. Introduction

Traffic-driven service provisioning using machine learning (ML), where network configurations are proactively adapted to accommodate future traffic demand, has demonstrated its effectiveness in network planning and operations by accurately modeling time-varying traffic behavior, leading to improved spectrum utilization, energy consumption, service disruptions, and capital and operational costs [1]–[6]. In essence, traffic prediction is a time series forecasting problem, where past traffic observations are used as input to predict future traffic demand, enabling proactive network optimization (i.e., predictive service provisioning) [1], [7]. In particular, deep neural networks (DNNs) [8] with recurrent units (i.e., long short-term memory (LSTM) units) [9], [10] have been widely adopted due to their competitive performance accuracies on real-world data. Alongside model accuracy, quantification of traffic prediction uncertainty (i.e., model and input uncertainty) and its consideration in proactive service provisioning was recently shown to be equally important, as even slight traffic inaccuracies may lead to quality-of-service (QoS) violations (i.e., service under-provisioning) and inefficient utilization of available resources (i.e., service over-provisioning) [11].

The vast majority of existing work on traffic-driven service provisioning has focused on the single-step ahead prediction problem (e.g., hourly predictions), with the predictions driving periodic service (re)provisioning [1]. In this context, prediction uncertainty has been considered in a few studies [6], [7], [11], [12], with the vast majority of existing work based on estimating prediction uncertainty through myopic margins, ignoring the fact that each input pattern is subject to different levels of uncertainty. More recently, deep-quantile inference [3], [13] has been adopted for estimating traffic prediction uncertainty, demonstrating the advantage of accurately quantifying and representing model and input uncertainty for network (re)optimization operations. Building upon these findings, work in [14] further demonstrated that Monte Carlo (MC) dropout inference offers an effective framework for uncertainty estimation, outperforming deep-quantile inference in both margin reduction and flexibility in efficiently tuning certainty levels according to targeted QoS requirements. Specifically, [11], [14] showed that MC dropout inference effectively tunes certainty levels by analyzing the predictive distribution obtained from a single trained DNN model. In contrast, quantile inference requires training multiple DNN models to estimate different quantiles (i.e., certainty levels), which significantly increases computational cost [14].

An open gap in the literature is the estimation of prediction uncertainty in multi-step ahead forecasting (i.e., predicting traffic for several hours ahead) and the effective exploitation of such estimates in a multi-period planning framework. Even though multi-step ahead prediction driving service provisioning has been previously studied [9], [10], [15], uncertainty estimation over predictions that span several hours ahead remains unexplored. In previous work [15] traffic-driven routing and spectrum allocation (RSA) heuristics exploiting multi-step ahead predictions were developed, demonstrating that predictive information over longer planning horizons is likely to improve service disruptions, while efficiently utilizing available spectrum resources. This work extends [15] by quantifying and representing uncertainty in multi-step ahead prediction, with this information subsequently used for proactive service provisioning. To this end, an uncertainty-aware traffic-driven service provisioning scheme for multi-period optical networks is proposed, further improving spectrum utilization, against a baseline, myopic, margin estimation approach. Additionally, it is shown that longer prediction horizons are likely to reduce service disruptions at the cost of increased over-provisioning.

II. Multi-step Network Traffic Prediction

An encoder-decoder LSTM (ED-LSTM) model, originally designed for sequence-to-sequence learning [16], is employed to perform multi-step-ahead traffic demand prediction. The objective of the ED-LSTM model is to learn a non-linear mapping $f(\cdot)$ that accurately forecasts future traffic demands (e.g., bit-rates) based on past and present traffic observations. In this work, $\mathbf{x}_t = [x_{t-r}, \cdots, x_{t-1}, x_t]$ denotes the vector of past and current traffic traces, while the corresponding future traffic demands are given by $\mathbf{y}_t = [y_{t+1}, y_{t+2}, \cdots, y_{t+u}]$, where r represents the number of past observations, t denotes the current time interval during which the traffic demand x_t is

H. Maryam, G. Savva, G. Ellinas, and T. Panayiotou are with the Department of Electrical and Computer Engineering and the KIOS Research and Innovation Center of Excellence (KIOS CoE), University of Cyprus, Nicosia, Cyprus. Emails: {maryam.hafsa, savva.giannis, gellinas, panayiotou.tania}@ucy.ac.cy

This work has been supported by the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy.

ISBN 978-3-903176-82-9

observed, and u is the number of future prediction steps (i.e., future planning intervals).

After training, the ED-LSTM model generates future traffic predictions according to $y_{t+u'}$= $f(\mathbf{x}_t, y_{t+1}, y_{t+2}, \ldots, y_{t+u'-1})$, where u' denotes a future time step such that $u' \leq u$. In previous work [15], the ED-LSTM model was trained using the Adam optimization algorithm to minimize the mean squared error (MSE) loss function, thereby targeting mean model responses in the form of point-estimate predictions. In this work, we additionally adopt Monte Carlo (MC) dropout inference to capture predictive uncertainty.

III. Modeling Uncertainty

Uncertainty is modeled in this work according to the MC dropout technique [17]. MC dropout is fundamentally based on approximating the posterior distribution over the weights of a DNN, given the observed data, which can be achieved by enabling dropout [18] during both training and inference. During inference, S stochastic forward passes are performed (i.e., using dropout) for each input traffic sequence $\mathbf{x}_t$, with each pass returning a u-dimensional prediction vector $\hat{\mathbf{y}}_t = \left[\hat{y}^{(s)}_{t+1}, \hat{y}^{(s)}_{t+2}, \ldots, \hat{y}^{(s)}_{t+u}\right]$, where $s \in 1, \ldots, S$. Hence, MC dropout enables multiple predictions to be made for each input traffic sequence $\mathbf{x}_t$ by sampling ED-LSTM parameters (i.e., weights) from the approximated posterior distribution (i.e., by enabling dropout multiple times during inference) to ultimately approximate predictive distribution, $p(\mathbf{y}_t|\mathbf{x}_t)$. Hence, unlike conventional ED-LSTMs, where a point-based traffic prediction is possible for a given input traffic sequence, MC dropout provides both a point-based traffic prediction, by averaging over the multiple predictions, and a measure of uncertainty obtained by appropriately analyzing the resulting predictive distribution, as described next.

IV. Obtaining Multi-Step Traffic Prediction Uncertainty

To derive a traffic prediction upper bound for each input traffic sequence $\mathbf{x}_t$, S forward passes are analyzed by computing the cumulative distribution function (CDF) of $p(\mathbf{y}_t|\mathbf{x}_t)$, for each prediction step u'. This enables the estimation of the upper bound $\hat{y}^q_{t+u'}$ for various certainty levels q. Particularly, the CDF is used to obtain the value $\hat{y}^q_{t+u'}$ such that $\Pr(y_{t+u'} \leq \hat{y}^q_{t+u'} \mid \mathbf{x}_t, y_{t+1}, y_{t+2}, ..., y_{t+u'-1})$=$q$, where $y_{t+u'}$ is the true traffic demand at $t+u'$ of sequence $\mathbf{x}_t$, and q is the predefined certainty level. Since the goal is to derive a prediction upper bound, the probability value q should be high enough to ensure that QoS violations (i.e., unserved traffic) will be at an acceptable threshold. Notably, multiple values of q may be considered in the presence of diverse QoS requirements. Specifically, to estimate $\hat{y}^q_{t+u'}$ for various certainty levels, S stochastic forward passes are exploited (i.e., $s = 1, ..., S$ with $\hat{y}^{(s)}_{t+u'}$ representing the s^{th} forward pass), returning $\hat{y}^q_{t+u'}$ according to Algorithm 1.

Given the upper traffic prediction estimates, $\hat{y}^q_{t+u'}$, for various q values, each q can be evaluated in terms of unserved traffic (i.e., resulting when the upper traffic prediction estimate is lower than the true traffic demand), and the q that satisfies the predefined QoS requirement can be selected accordingly.

Algorithm 1 Estimation of $\hat{y}^q_{t+u'}$.

for $q = 0.5 : l$ **do** (where $l \leq 0.95$)
  for $s = 1 : S$ **do**
    $Pr(y_{t+u'} \leq \hat{y}^{(s)}_{t+u'} \mid \mathbf{x}_t, y_{t+1}, y_{t+2}, ..., y_{t+u'-1})$ = $\frac{1}{S}\sum_{s'=1}^{S} r_{s'}$,
  where $r_{s'} = 1$ if $\hat{y}^{(s')}_{t+u'} \leq \hat{y}^{(s)}_{t+u'}$, else $r_{s'} = 0$
    if $Pr(y_{t+u'} \leq \hat{y}^{(s)}_{t+u'} \mid \mathbf{x}_t, y_{t+1}, y_{t+2}, ..., y_{t+u'-1}) \approx q$ **then**
      break
    end if
  end for
  $\hat{y}^q_{t+u'} = \hat{y}^{(s)}_{t+u'}$
end for

Evaluation of each certainty level q is, in essence, carried out on a labeled test dataset. Then, the most suitable q can be used to offline (re)optimize network resources for future planning intervals (by considering that future traffic demand of established lightpaths may fluctuate according to $\hat{y}^q_{t+u'}$).

V. Dataset Formulation and Generation

Training dataset D is derived from real traffic traces of the Abilene backbone network (http://sndlib.zib.de/home.action), which consists of 12 nodes and 30 links. The dataset provides bit-rate information (in Mbps) for each node pair, recorded at 5-minute intervals over a 6-month period in the form of traffic demand matrices. For each source–destination pair (k, k'), a separate dataset D_k is constructed, where the destination k' is randomly selected from the Abilene network topology. Specifically, the traffic patterns in each D_k represent the aggregated bit-rate originating from source node k over successive 5-minute intervals across the entire Abilene dataset.

Given the aggregated bit rates, the input traffic patterns for each D_k are constructed as $\mathbf{x}_t = [x_{t-r}, \ldots, x_{t-1}, x_t]$, where $x_{t-i} = [f^1_{t-i}, f^2_{t-i}, \ldots, f^\tau_{t-i}], \quad \forall i = 0, 1, \ldots, r, \ \forall t = 1, 2, \ldots, T$. Here, x_t is a vector consisting of τ traffic fluctuations within time interval t. The ground truths of each D_k are defined as $\mathbf{y}_t = [y_{t+1}, y_{t+2}, \ldots, y_{t+u}]$, where $y_{t+j} = \max\{f^1_{t-j}, f^2_{t-j}, \ldots, f^\tau_{t-j}\}, \quad \forall t = 1, 2, \ldots, T, \ \forall j = 1, 2, \ldots, u$. This formulation targets the prediction of the maximum traffic demand that may occur in interval $t+j$, thereby mitigating potential QoS violations.

As the original Abilene dataset provides bit-rate information at 5-minute granularity, it is assumed that traffic fluctuations in all D_k occur on the same time scale. Specifically, each time interval t spans 30 minutes, resulting in τ=6 traffic fluctuations within each time interval.

VI. Model Training and Evaluation

For model training and evaluation, each dataset D_k consists of T=3000 sequential traffic patterns generated using a sliding-window approach, corresponding to the first 17 days of the original dataset. All datasets are normalized to lie within the range $[0, 1]$. For model training and validation, 80% of the traffic patterns in D_k are used (i.e., the traffic patterns spanning the first 13 days of D_k), while the remaining 20% are reserved for model evaluation.

For each source node k, multiple ED-LSTM models are trained and optimized using the Adam optimizer under different hyperparameter configurations (e.g., learning rate, batch

size, delta value, hidden units, dropout rates), in order to identify the best-performing ED-LSTM model configuration. Specifically, all ED-LSTM models are trained for 500 epochs using the following hyperparameter options: (i) learning rates $\lambda \in \{0.01, 0.001, 0.0001\}$; (ii) batch sizes $\beta \in \{32, 64, 128\}$; (iii) two hidden layers with the number of hidden units $h \in \{32, 64, 128\}$; (iv) the Huber loss function with delta values $\delta \in \{0.5, 1.0\}$; and (v) dropout rates $P \in \{0.1, 0.2, 0.3\}$. This section reports the best-performing configurations for each model, selected after extensive hyperparameter tuning. Note that, in this work the Huber loss function is opted mainly due to its robustness to outliers. In general, this function behaves like the MSE for small prediction errors, promoting stable and accurate fitting, and like the mean absolute error (MAE) for large errors, thereby limiting the influence of extreme deviations. This hybrid behavior prevents anomalous traffic spikes from disproportionately affecting model training, resulting in more stable convergence and improved generalization, particularly for network traffic demand data. Furthermore, an early stopping criterion is employed to monitor the validation loss and terminate training if no improvement is observed for 20 consecutive epochs.

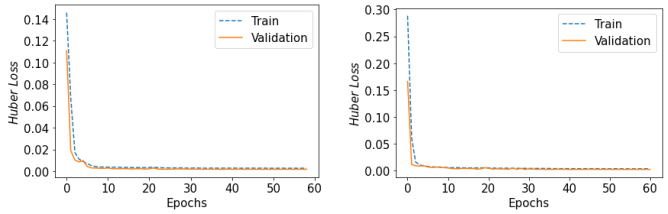

Fig. 1: ED-LSTM training evolution across the STTLng (k=11) Abilene node using the MC dropout technique for u=4 (left) and u=6 (right).

Figure 1 illustrates the ED-LSTM training evolution using D_{11} for $u = 4, 6$ prediction steps. Clearly, the mean training and test loss converge to a value close to zero in both models. In Fig. 1, a mean test loss value of 0.00411 and 0.00461 is obtained for u=4 and u=6, respectively. Interestingly, as u increases, the loss values of the models also increase. This behavior is expected because longer planning intervals introduce increased prediction complexity and uncertainty, making the problem harder. It should be noted that a similar training and convergence behavior is observed across all D_k datasets (not illustrated due to space limitations).

Regarding the mean response values (i.e., point estimates) for each input pattern $\mathbf{x}_t$, these are computed by averaging the S=100 predictions obtained through the MC dropout stochastic passes during inference. Specifically, mean response values are computed as $\hat{\psi}_{t+i} = \frac{1}{S}\sum_s \hat{y}^{(s)}_{t+i}$, for all $i = 1, .., u$ prediction steps. Illustrative examples of the predictive distributions, formed by the S stochastic passes, are shown in Fig. 2. Specifically, the figure illustrates the predictive distribution of D_{11} for the $t+1$, $t+2$, $t+3$, and $t+4$ prediction steps (from left to right), given test traffic sequence $\mathbf{x}_t$ (where traffic sequence $\mathbf{x}_t$ is randomly chosen from the test dataset). The green lines represent the ground truth, the red lines denote mean response values $\hat{\psi}_{t+i}$, and uncertainty is visualized by the bell-shaped distributions formed by the S stochastic passes (i.e., the $\hat{y}^{(s)}_{t+i}$ values). Overall, Fig. 2 shows that the magnitude of uncertainty around the mean response values varies across prediction steps. Hence, both the input patterns and the prediction horizon length are subject to different levels of uncertainty.

VII. Baseline Margin Estimation and Comparisons

Margins are typically introduced to compensate for prediction uncertainty. Two baseline margin estimation schemes are widely used in the state-of-the-art, typically based either on the maximum error that is observed in the test dataset or on the empirical rule error [7], [11], [14]. This work adopts as a baseline the empirical rule error margin that is shown to yield better results than the maximum error-based margin [11]. Specifically, $M_{e,i}$ denotes the empirical rule error margin across the predictions at step i, computed as the mean error value, μ_i, plus three standard deviations, σ_i, for all $i = 1, ..., u$. Specifically, $M_{e,i} = \mu_i + 3\sigma_i$, with $\mu_i = \frac{1}{(T'/u)} \sum\limits_{\substack{t=0,\ s.t. \\ t\ mod\ u=0}}^{T'} (y_{t+i} - \hat{\psi}_{t+i})$, and $\sigma_i^2 = \frac{1}{(T'/u)} \sum\limits_{\substack{t=0,\ s.t. \\ t\ mod\ u=0}}^{T'} (y_{t+i} - \hat{\psi}_{t+i} - \mu_i)^2$, where T' is the number of traffic sequences in the test dataset.

Clearly, unlike MC-based uncertainty treatment (Alg. 1), empirical rule error compensates for prediction uncertainty by shifting all predictions by $M_{e,i}$. Specifically, all $\hat{\psi}_{t+i}$ predictions are increased by $M_{e,i}$ prior to service provisioning to mitigate QoS violations. Hence, to enable comparison with the baseline, an equivalent average MC margin is computed, denoted as $\bar{M}^q_{mc,i}$, given certainty threshold q. In particular, MC margins are computed dynamically by analyzing the predictive distribution for each input sequence $\mathbf{x}_t$ for all prediction steps $i = 1, .., u$, according to:

$$\bar{M}^q_{mc,i} = \frac{1}{(T'/u)} \sum_{\substack{t=0,\ s.t. \\ t\ mod\ u=0}}^{T'} \left(\hat{y}^q_{t+i} - \hat{\psi}_{t+i}\right), \qquad (1)$$

where $\hat{\psi}_{t+i}$ is the mean response of the MC dropout model for the given test pattern and $\hat{y}^q_{t+i}$ is the upper bound traffic estimate derived by the MC dropout model for certainty level q (Alg. 1). Also, since each prediction step corresponds to a different $M^q_{mc,i}$ value, the average equivalent margin is computed by averaging over all traffic sequences $t \in \{1, 2, \ldots, T'\}$ for each $i \in \{1, 2, \ldots, u\}$, instead of just computing $M^q_{mc,i}$ for each $t \in \{1, 2, \ldots, T'\}$.

Tables I and II summarize the margin values obtained for all schemes, including baseline margin $M_{e,i}$ and equivalent MC margin $\bar{M}^q_{mc,i}$, evaluated for two certainty thresholds, $q = 0.90$ and $q = 0.95$. According to these results, $\bar{M}^q_{mc,i}$ consistently outperforms the baseline $M_{e,i}$ margin for both certainty thresholds q considered. Interestingly, as q increases, the corresponding $\bar{M}^q_{mc,i}$ values also increase, since higher q values reduce the percentage of traffic sequences for which their true traffic demand exceeds the predicted upper-bound estimates $\hat{y}^q_{t+i}$. Furthermore, margin values tend to be higher for u=6 (i.e., the longest prediction horizon considered -

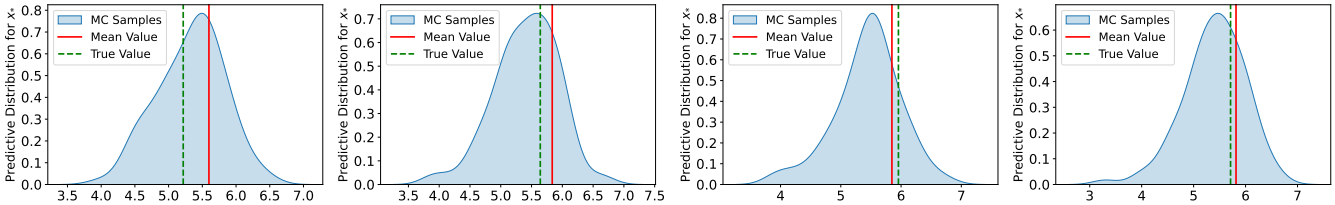

Fig. 2: Predictive distributions for different test input sequences x_* from dataset D_{11} over u=4 planning intervals. Specifically, the figure shows the estimated probability density function (y-axis) of the Monte Carlo (MC) estimates (x-axis) for various input values.

Table II) as compared to u=4 (Table I), reflecting the higher uncertainty associated with longer prediction horizons.

Overall, $\bar{M}^q_{mc,i}$ achieves margin reductions ranging between 29% to 98% for u=4 and between 9% to 97% for u=6, when compared to the baseline $M_{e,i}$ scheme, depending on the model and the selected q value considered. On average, margin reduction improvements of 82%(78%) are observed for u=4, and 73%(66%) for u=6 for $q = 0.90(0.95)$, compared to the baseline $M_{e,i}$ scheme. These results indicate the importance of appropriately quantifying uncertainty, as opposed to the myopic, empirical, margin scheme. To evaluate the impact of margin reduction on service provisioning (i.e., QoS violations, over-provisioning, disruptions) a multi-period planning elastic optical network (EON) is utilized for both uncertainty estimation frameworks considered.

VIII. Uncertainty-Aware Multi-Period RSA

Several spectrum allocation (SA) heuristics are developed to exploit multi-step ahead predictions and uncertainty information, aiming to effectively utilize available spectrum resources. For all SA heuristics, the RSA is solved by first solving the routing sub-problem according to the κ-shortest path algorithm (with $\kappa = 3$), followed by a first-fit SA strategy. A connection is blocked if a feasible SA cannot be found for any candidate route. The various SA schemes developed, therefore, differ in the way the selected spectrum size is chosen for each connection request and future planning interval. Specifically, following [15], two heuristic algorithms are considered and extended to additionally leverage uncertainty information; that is, the multi-step maximum demand spectrum allocation (MMD-SA) and the multi-step average demand spectrum allocation (MAD-SA) heuristics. Note that since a prediction interval $t+i$ is equivalent to a planning interval $t+i$ (i.e., an interval where RSA may be applied, possibly resulting to dynamically reconfiguring a connection), these two terms are used interchangeably. To this end, heuristic descriptions follow:

• **MMD-SA with Empirical Margin**: For each connection k, the maximum bit-rate predicted across all $t+1,..,t+u$ intervals is selected and increased by the corresponding empirical margin prior to SA. Specifically, for each connection k, the selected bit-rate is identified and computed as $y^k_v = \max\{\hat{\psi}^k_{t+i}\}^u_{i=1} + M^k_{e,v}$, where $v = \arg\max_v\{\hat{\psi}^k_{t+v}\}^u_{v=1}$.

• **MMD-SA with MC Margin given q**: For each connection k, the maximum bit-rate predicted across all $t+1,..,t+u$ intervals, by means of MC dropout inference given a certainty threshold q, is selected. Specifically, for each connection k, the selected bit-rate is identified and computed as $y^k_v = \max\{\hat{y}^{q,k}_{t+i}\}^u_{i=1}$, where $\hat{y}^{q,k}_{t+i}$ is given by Alg. 1.

• **MAD-SA with Empirical Margin**: For all connection requests, the bit-rates selected and increased by the corresponding empirical margin prior to SA, are those corresponding to the predictions generated at v, with v corresponding to the most loaded planning interval across all connections (i.e., the interval where the aggregated bit-rate is expected to be the highest). Specifically, $v = \arg\max_v\{\sum_k \hat{\psi}^k_{t+v}\}^u_{v=1}$, and $y^k_v = \psi^k_{t+v} + M^k_{e,v}$ is the bit-rate to be allocated to serve connection k.

• **MAD-SA with MC Margin given q**: For all connection requests, the bit-rates selected are those corresponding to the predictions generated at τ, by means of MC dropout inference, given a certainty threshold q, with v corresponding to the most loaded planning interval across all connections (i.e., the interval where the aggregated bit-rate is expected to be the highest). Specifically, $v = \arg\max_v\{\sum_k \hat{y}^{q,k}_{t+v}\}^u_{v=1}$, where $\hat{y}^{q,k}_{t+v}$ is given by Alg. 1, and $y^k_v = \hat{y}^{q,k}_{t+v}$ is the bit-rate to be allocated to connection k.

For all heuristics, the selected bit-rates are chosen to accommodate the traffic demand for all connections and for all u future time intervals. Hence, a SA heuristic is activated every u time intervals; that is, at every t where $t \mod u$=0. As an example, given that u=4, a SA heuristic is activated at t=0 to accommodate the expected future traffic demand at t=1, 2, 3, 4, at t=4 to accommodate the expected future demand at t=5, 6, 7, 8, and so on. The selected bit-rates are then converted into the required number of frequency slots (FSs), following a conventional distance-adaptive modulation scheme, utilizing BPSK, QPSK, 8-QAM, and 16-QAM modulation formats [19] and considering the distance of the precomputed κ-shortest paths.

In general, activation of a SA heuristic may lead to connection reconfigurations, possibly causing service disruptions and QoS violations. To avoid unnecessary service disruptions, the following spectrum reduction-expansion [20] and re-allocation policies are followed: (i) *Spectrum Reduction*: if the allocated spectrum size exceeds the spectrum size to be allocated for the next time intervals, then a spectrum reduction policy is applied (which is always feasible); (ii) *Spectrum Expansion*: if the allocated spectrum size is less than the spectrum size to be allocated for the next time intervals, first, an in-place spectrum adjustment (i.e., allocating additional adjacent FSs along the same route) is attempted to minimize service disruption. (iii) *Spectrum Re-allocation*: In case expansion is not feasible (i.e., additional adjacent FSs are not available), the connection is

TABLE I: Margin comparison (in Gbps) for u=4 prediction steps.

k		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$M_{e,i}$	$M_{e,1}$	0.7246	30.3411	0.7130	0.1936	0.8560	0.3236	0.8870	1.7595	0.4977	0.4664	2.1619	0.0310
	$M_{e,2}$	0.8205	37.3187	0.9519	0.3507	1.2934	0.4364	2.7884	2.2761	0.3834	0.6661	1.9189	0.0418
	$M_{e,3}$	1.0903	46.2998	1.0453	0.7115	0.9606	0.4620	10.2311	3.7501	0.3815	0.8427	1.9504	0.0464
	$M_{e,4}$	1.2315	40.9655	1.2848	0.5666	1.0148	0.4472	2.4776	3.0863	0.5123	0.7951	2.4899	0.0441
$\tilde{M}^{0.9}_{mc,i} \pm std$ ($\times 10^{-2}$)	$M_{mc,1}$	7.38 ± 2.81	116.05 ± 123.64	16.54 ± 6.85	11.16 ± 1.28	14.26 ± 2.51	7.79 ± 1.77	51.08 ± 16.00	44.18 ± 13.56	10.27 ± 3.71	10.82 ± 4.16	54.11 ± 8.43	0.466 ± 0.126
	$M_{mc,2}$	7.66 ± 2.53	105.28 ± 110.74	14.30 ± 4.48	12.54 ± 1.75	14.21 ± 2.49	7.99 ± 1.23	42.21 ± 10.92	48.26 ± 8.56	10.76 ± 3.07	11.03 ± 3.00	46.06 ± 6.81	0.485 ± 0.093
	$M_{mc,3}$	7.49 ± 2.39	103.55 ± 113.93	14.19 ± 3.95	12.84 ± 1.89	16.18 ± 4.16	7.98 ± 1.22	38.93 ± 10.69	50.47 ± 8.64	10.25 ± 2.66	10.30 ± 2.53	46.11 ± 7.61	0.687 ± 0.088
	$M_{mc,4}$	7.39 ± 2.15	102.01 ± 116.50	14.33 ± 3.82	13.35 ± 2.08	17.16 ± 3.80	7.73 ± 1.12	37.18 ± 9.75	50.50 ± 7.49	9.98 ± 2.55	10.54 ± 2.45	47.33 ± 8.92	0.654 ± 0.075
$\tilde{M}^{0.95}_{mc,i} \pm std$ ($\times 10^{-2}$)	$M_{mc,1}$	9.64 ± 3.91	149.51 ± 170.74	21.12 ± 8.96	13.81 ± 1.80	17.71 ± 3.36	10.03 ± 2.34	60.79 ± 18.80	51.95 ± 17.44	12.62 ± 4.48	12.91 ± 4.76	66.59 ± 9.97	0.574 ± 0.145
	$M_{mc,2}$	9.67 ± 3.03	131.69 ± 135.20	18.00 ± 5.51	15.47 ± 2.20	17.69 ± 2.88	10.27 ± 1.60	51.27 ± 13.90	56.19 ± 10.31	13.31 ± 3.83	13.00 ± 3.40	56.71 ± 8.46	0.604 ± 0.109
	$M_{mc,3}$	9.55 ± 3.10	128.31 ± 141.20	18.17 ± 5.25	15.86 ± 2.36	20.10 ± 5.22	10.24 ± 1.61	47.32 ± 12.09	59.30 ± 11.07	12.71 ± 3.22	12.19 ± 2.84	56.74 ± 9.41	0.848 ± 0.112
	$M_{mc,4}$	9.43 ± 2.71	129.05 ± 143.38	18.04 ± 4.80	16.46 ± 2.52	21.76 ± 4.91	9.90 ± 1.44	45.89 ± 11.73	59.23 ± 9.63	12.36 ± 3.08	12.57 ± 2.96	59.45 ± 10.80	0.802 ± 0.104

TABLE II: Margin comparison (in Gbps) for u=6 prediction steps.

k		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$M_{e,i}$	$M_{e,1}$	0.9283	31.4270	1.1789	0.3886	0.7105	0.3851	3.1211	3.2994	0.5628	0.4673	1.6924	0.0438
	$M_{e,2}$	0.9227	33.1319	1.2996	0.3419	1.2425	0.4886	12.6915	2.8934	0.3840	0.6277	1.8474	0.0324
	$M_{e,3}$	0.8930	37.9983	0.9513	0.2297	0.9087	0.5153	2.3101	2.2888	0.3710	0.7206	2.1793	0.0276
	$M_{e,4}$	0.9929	45.7202	1.1656	0.2732	0.8128	0.4897	1.9143	2.6747	0.4246	0.8882	2.6153	0.0363
	$M_{e,5}$	1.3159	49.4844	1.3517	0.3123	1.1221	0.5525	1.9354	3.4268	0.4466	0.9479	2.9468	0.0290
	$M_{e,6}$	1.5333	51.2465	1.2585	0.6780	1.3010	0.5284	2.3177	3.4290	0.5942	0.8355	2.7243	0.0341
$\tilde{M}^{0.90}_{mc,i} \pm std$ ($\times 10^{-2}$)	$M_{mc,1}$	21.31 ± 7.44	200.60 ± 236.05	25.41 ± 6.62	16.65 ± 2.05	19.81 ± 6.41	12.51 ± 2.61	109.56 ± 31.20	60.07 ± 20.28	16.05 ± 4.23	16.64 ± 2.25	80.12 ± 9.29	0.766 ± 0.145
	$M_{mc,2}$	20.87 ± 4.68	179.86 ± 221.37	21.72 ± 5.47	18.04 ± 2.13	23.74 ± 3.27	10.42 ± 1.32	120.60 ± 29.37	65.69 ± 10.29	16.56 ± 3.48	19.13 ± 1.65	69.57 ± 8.64	0.784 ± 0.144
	$M_{mc,3}$	21.29 ± 4.50	169.51 ± 201.38	20.04 ± 4.98	16.91 ± 2.07	22.38 ± 4.23	10.01 ± 1.28	113.10 ± 24.48	65.34 ± 10.07	15.71 ± 2.87	18.07 ± 1.86	67.88 ± 8.23	0.880 ± 0.129
	$M_{mc,4}$	21.40 ± 4.70	155.09 ± 190.88	19.17 ± 5.10	16.12 ± 1.77	24.14 ± 2.84	9.83 ± 1.19	120.17 ± 25.04	63.90 ± 10.68	15.41 ± 3.07	18.26 ± 1.59	64.04 ± 8.40	1.156 ± 0.217
	$M_{mc,5}$	20.97 ± 3.83	153.11 ± 191.81	18.59 ± 4.49	16.76 ± 1.85	20.75 ± 5.15	9.80 ± 1.10	135.06 ± 29.88	62.54 ± 9.82	15.25 ± 2.75	17.26 ± 1.37	62.49 ± 8.38	0.851 ± 0.170
	$M_{mc,6}$	21.08 ± 3.76	147.66 ± 183.55	17.30 ± 4.15	16.63 ± 1.80	22.94 ± 3.17	9.87 ± 1.10	145.10 ± 31.93	62.39 ± 10.01	15.24 ± 2.43	18.34 ± 1.73	61.88 ± 8.68	1.039 ± 0.117
$\tilde{M}^{0.95}_{mc,i} \pm std$ ($\times 10^{-2}$)	$M_{mc,1}$	26.17 ± 8.97	251.50 ± 290.80	31.97 ± 8.67	20.72 ± 2.55	24.32 ± 6.96	15.44 ± 3.20	142.74 ± 44.81	75.88 ± 29.37	20.79 ± 5.72	20.83 ± 2.79	99.13 ± 11.59	0.943 ± 0.167
	$M_{mc,2}$	26.05 ± 6.23	223.22 ± 271.21	27.29 ± 7.13	22.40 ± 2.49	28.85 ± 3.11	12.84 ± 1.64	156.06 ± 44.76	82.14 ± 13.70	21.02 ± 4.35	23.84 ± 2.20	86.60 ± 12.13	0.966 ± 0.174
	$M_{mc,3}$	25.87 ± 6.17	212.10 ± 250.62	25.10 ± 6.28	20.96 ± 2.19	27.54 ± 4.25	12.35 ± 1.52	143.50 ± 36.44	81.93 ± 13.52	19.78 ± 3.28	22.57 ± 2.35	82.75 ± 9.91	1.098 ± 0.145
	$M_{mc,4}$	26.47 ± 6.39	191.94 ± 231.11	23.93 ± 5.92	19.93 ± 2.08	29.53 ± 2.84	12.15 ± 1.45	151.38 ± 35.52	80.04 ± 12.72	19.28 ± 3.54	22.69 ± 2.11	78.51 ± 9.73	1.417 ± 0.277
	$M_{mc,5}$	25.64 ± 4.79	192.34 ± 234.77	23.24 ± 5.20	20.70 ± 2.19	25.76 ± 5.54	12.15 ± 1.40	173.71 ± 41.81	78.78 ± 12.49	19.21 ± 3.37	21.37 ± 1.89	76.30 ± 9.50	1.053 ± 0.186
	$M_{mc,6}$	25.93 ± 4.86	183.79 ± 221.96	21.73 ± 5.40	20.89 ± 2.33	28.33 ± 3.58	12.30 ± 1.37	183.83 ± 43.99	78.33 ± 12.91	19.16 ± 3.23	22.76 ± 2.17	75.50 ± 10.49	1.294 ± 0.158

re-allocated (i.e., rerouted and/or reassigned new spectrum resources), leading to a disruption. If re-allocation is not possible, then the connection is blocked.

IX. Performance Evaluation of Uncertainty-Aware Multi-Period RSA

All uncertainty-aware multi-period RSA schemes developed are evaluated and compared considering the Abilene network topology (described in Section V) with each fiber link consisting of 320 FSs with a channel spacing of 12.5 GHz and a symbol rate (R_s) of 10.7 Gbaud. For the predictions and uncertainty estimates, the models trained in Section VI with the corresponding test datasets are considered, with each model corresponding to a distinct source-destination pair, k-k', where the destination node k' is randomly selected from the Abilene network topology. For all schemes, both predicted and true bit rates are scaled by a factor of 30 to align the spectrum demand with the available network capacity.

Table III summarizes the results for u=4, 6 prediction steps and for all SA schemes developed, in terms of: (i) number of blocked connections, (ii) number of connection disruptions, (iii) average underprovisioning measured in both Gbps and FSs, and (iv) average overprovisioning measured in both Gbps and FSs. While the events of a blocked connection and a connection disruption are described in Section VIII, underprovisioning and overprovisioning are in general computed by comparing the SAs selected by the SA heuristic under examination, with the true traffic demands. Specifically, SAs are evaluated against the true fluctuations (i.e., against the true traffic fluctuations in $\mathbf{y_t} = [\mathbf{y_{t+1}}, .., \mathbf{y_{t+u}}]$) in the test datasets D_k, resulting in underprovisioning when the true traffic fluctuation is above the bandwidth allocated and overprovisioning when the true traffic fluctuation is below the bandwidth allocated. Both total underprovisioning and overprovisioning are averaged over all planning intervals (i.e., SA heuristic activations), connections, and fluctuations.

According to the results, all schemes result in zero blocking. Additionally, SA heuristics (MMD-SA and MAD-SA) with MC margin significantly and consistently outperform SA schemes with empirical margin in terms of overprovisioning. This indicates higher spectrum savings for both certainty thresholds (i.e., q=0.90, 0.95) and both planning intervals, u=4, 6 considered. Specifically, MMD-SA with MC margin achieves spectrum savings of up to 79%(77%) in Gbps compared to MMD-SA with empirical margin M_e when u=4(6). A similar trend is observed for the MAD-SA with MC margin scheme. Specifically, spectrum savings of up to 79%(75%) in Gbps are achieved compared to the MAD-SA with empirical margin when u=4(6). Compared to each other, the two certainty thresholds, q, considered perform similarly in terms of spectrum savings ($q = 0.95$ slightly increases overprovisioning compared to $q = 0.9$ - i.e., up to 5%- to provide a higher a certainty level). Additionally, it is observed that as the planning interval increases (i.e., as u increases), overprovisioning also increases. This is to be expected, as spectrum resources are allocated to accommodate the expected traffic demand for a longer future horizon, inevitably increasing the overall allocated resources.

Regarding underprovisioning (i.e., unserved traffic), Table III shows that both SA heuristics (MMD-SA and MAD-SA) exhibit a similar trend across the evaluated scenarios. The SA schemes with empirical margins result in negligible unserved traffic across all planning intervals, while the SA schemes with MC margins exhibits a slightly higher amount of unserved traffic. However, in the latter case, this increase in unserved traffic remains very small, while improved spectrum efficiency is achieved (as the unutilized traffic is greatly reduced), indicating a clear trade-off between spectrum savings and the amount of unserved traffic/QoS.

TABLE III: Evaluation of different spectrum allocation schemes for u=4 and u=6 prediction steps.

		MMD-SA			MAD-SA		
		Empirical Margin, M_e	MC Margin, $q = 0.90$	MC Margin, $q = 0.95$	Empirical Margin, M_e	MC Margin, $q = 0.90$	MC Margin, $q = 0.9$
$u = 4$	Blocked Connections	0	0	0	0	0	0
	Disruptions	8	10	11	10	8	9
	Avg. Underprovisioning (in Gbps)	0.160	3.741	3.583	0.163	3.779	3.639
	Avg. Overprovisioning (in Gbps)	199.256	41.529	43.303	196.093	40.035	41.744
	Avg. Underprovisioning (in FSs)	0.006	0.107	0.101	0.007	0.111	0.106
	Avg. Overprovisioning (in FSs)	6.030	1.044	1.117	5.912	0.978	1.047
$u = 6$	Blocked Connections	0	0	0	0	0	0
	Disruptions	7	9	8	7	10	9
	Avg. Underprovisioning (in Gbps)	0.128	3.209	3.112	0.169	3.242	3.130
	Avg. Overprovisioning (in Gbps)	225.016	50.467	54.179	197.253	48.788	52.323
	Avg. Underprovisioning (in FSs)	0.006	0.096	0.090	0.009	0.099	0.093
	Avg. Overprovisioning (in FSs)	6.804	1.350	1.508	5.513	1.280	1.426

Regarding service disruptions, SA schemes with MC margins tend to experience a higher number of disruptions compared to the SA schemes with empirical margin schemes, mainly because empirical margins tend to allocate more resources compared to the MC case, leading to a reduced number of reallocated connections (i.e., there exists an overprovisioning-disruptions trade-off). Note that for the longer planning interval examined (i.e., u=6), the number of disruptions decreases for the vast majority of scenarios examined. This reduction can be attributed to the availability of longer temporal information, which will likely lead to improved SA decisions, and to less frequent (re)allocations that the connections undergo. This, however, greatly depends on the quality of the predictions and the capability of the model to accurate predict traffic for longer planning intervals (i.e., when the planning horizon becomes very long, model accuracy may degrade, possibly leading to service disruption due to inaccurate SA decisions).

X. Conclusion

This work demonstrates the importance of appropriately representing uncertainty over multi-step ahead traffic predictions, towards effective SA decisions. Specifically, the results demonstrate the ability of MC drop-out inference to more efficiently utilize available spectrum resources compared to the traditional, myopic, empirical margins. This is shown considering two uncertainty-aware SA heuristics (MMD-SA and MAD-SA), that differently exploit multi-step ahead traffic predictions. Overall, it is shown that when uncertainty is appropriately quantified and considered prior to SA decisions, significant spectrum savings can be achieved by both MMD-SA and MAD-SA heuristics; that is, MC dropout-based margins lead to SAs more efficiently utilizing the available network resources compared to empirical-based margins. Additionally it is shown that longer planning horizon are likely to lead to reduced service disruptions, provided that the predictions are of sufficient accuracy. Overall, the selection of a suitable SA scheme greatly depends on the targeted QoS requirements (i.e., tolerance in unserved traffic and/or service disruptions that may lead to traffic loss).

References

[1] T. Panayiotou *et al.*, "Survey on machine learning for traffic-driven service provisioning in optical networks," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 2, pp. 1412–1443, 2023.

[2] T. Panayiotou, S. Chatzis *et al.*, "A data-driven bandwidth allocation framework with QoS considerations for EONs," *IEEE/OSA J. Light. Technol.*, vol. 37, no. 9, pp. 1853–1864, 2019.

[3] T. Panayiotou and G. Ellinas, "Addressing traffic prediction uncertainty in multi-period planning optical networks," in *Proc. IEEE/OSA OFC*, 2022, pp. 1–3.

[4] S. Shakya *et al.*, "Minimize sub-carrier reallocation in elastic optical path networks using traffic prediction," in *Proc. IEEE GLOBECOM*, 2013, pp. 2352–2357.

[5] A. Coiro *et al.*, "Reducing power consumption in wavelength routed networks by selective switch off of optical links," *IEEE J. Sel. Top. Quantum Electron.*, vol. 17, no. 2, pp. 428–436, 2010.

[6] R. Alvizu *et al.*, "Energy efficient dynamic optical routing for mobile metro-core networks under tidal traffic patterns," *IEEE/OSA J. Light. Technol.*, vol. 35, no. 2, pp. 325–333, 2017.

[7] S. Troia *et al.*, "Deep learning-based traffic prediction for network optimization," in *Proc. IEEE ICTON*, 2018, pp. 1–4.

[8] X. Chen *et al.*, "Knowledge-based autonomous service provisioning in multi-domain elastic optical networks," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 152–158, 2018.

[9] M. Balanici and S. Pachnicke, "Classification and forecasting of real-time server traffic flows employing long short-term memory for hybrid E/O data center networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 13, no. 5, pp. 85–93, 2021.

[10] S. K. Singh *et al.*, "Machine-learning-aided dynamic reconfiguration in optical DC/HPC networks," in *Proc. ONDM*, 2022, pp. 1–6.

[11] H. Maryam *et al.*, "Uncertainty quantification and consideration in ML-aided traffic-driven service provisioning," *Computer Communications*, vol. 202, pp. 13–22, 2023.

[12] T. Panayiotou *et al.*, "Deep quantile regression for QoT inference and confident decision making," in *Proc. IEEE ISCC*, 2021, pp. 1–6.

[13] H. Maryam *et al.*, "Learning quantile QoT models to address uncertainty over unseen lightpaths," *Computer Networks*, vol. 212, pp. 1–10, 2022.

[14] H. Maryam, T. Panayiotou *et al.*, "Representing uncertainty in deep QoT models," in *Proc. IEEE MedComNet*, 2022, pp. 113–121.

[15] H. Maryam, T. Panayiotou, and G. Ellinas, "Multi-step traffic prediction for multi-period planning in optical networks," in *Proc. IEEE ICTON*, 2024, pp. 1–5.

[16] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv:1406.1078*, 2014.

[17] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.

[18] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[19] M. Jinno *et al.*, "Distance-adaptive spectrum resource allocation in spectrum-sliced elastic optical path network," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 138–145, 2010.

[20] K. Christodoulopoulos *et al.*, "Time-varying spectrum allocation policies and blocking analysis in flexible optical networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 1, pp. 13–25, 2013.