Projet ANR PICS -INRAE Oct. Prévision Immédiate Intégrée des Impacts des Crues Soudaines LSTM output correction of the conceptual rainfall-runo model GRD

Reyhaneh Hashemi, Pierre Javelle

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Projet ANR PICS – INRAE Oct. 2021

Prévision Immédiate Intégrée des Impacts des Crues Soudaines

LSTM output correction of the conceptual rainfall-runoff model GRD

Reyhaneh Hashemi

Contents

List of Figures

List of Tables

1 INTRODUCTION

Simulation of surface runoff by conceptual rainfall-runoff (R-R) models involves error due to uncertainties related to input data, hydrological uncertainties associated with the model structure, and parameter calibration. In a flood forecasting system the inherent impreciseness of R-R models become more problematic since the system performance relies highly on performances of the incorporated R-R model. One way to deal with this problem is to try to improve R-R prediction accuracy by correcting their simulations. This is equivalent to precisely predicting R-R simulation errors, i.e. the discrepancy between observed streamflow values and values computed by R-R models. The error prediction approach, also called error assimilation, offers several key advantages including that it is simple, fully automated, and applicable to any R-R model independent of its conceptual description and degree of complexity. In this approach, a supplementary stochastic model is first calibrated on the historical error data and is then used to predict future error. The predicted error is subsequently added as a correction term to the initial outputs of the R-R model. There are different types of stochastic error forecasting methods. Reviews agree on the following classes as the most popular ones.

An essential classical category of these methods is based on the fact that there is a persistence in R-R model output errors and they are mainly autocorrelated. Examples of this class include Auto-Regressive (AR), Auto-Regressive Moving Average (ARMA), and Auto-Regressive Integrated Moving Average (ARIMA) models (Schreider et al., 1997, Shamseldin et O'CONNOR, 1999, Toth et al., 1999, Goswami et O'Connor, 2010, Sun et al., 2017). Similarity based models are another type of error forecasting models. In these models, instances with similar outputs are considered to have similar errors. Time steps similar to a given time step are detected through the K Nearest Neighbor (KNN) algorithm and using a set of reference error-calibration data (Akbari et Afshar, 2013). Other classes include Dual Pass (Pagano et al., 2011, Liu et al., 2019), Bayesian Joint Probability (BJP) (Pokhrel et al., 2013), and hybrid methods combining, for instance, Kalman Filters (KFs) with ARMA (Bidwell et Griffiths, 1994) or Artificial Neural Networks (ANNs) (Muluye, 2011), or an ensemble based maximum a posteriori (MAP) estimation approach with a lag-aware ensemble Kalman smoother (EnKS) (Li et al., 2015), or Least Squares-Support Vector Machines (LS-SVM) with a 4D copula function (Liu et al., 2021), and so forth.

Following the purely ANN output assimilation models (Babovic et al., 2001, Anctil et al., 2003, Humphrey et al., 2016, Ghaith et al., 2020), it is only very recently that the Deep Learning LSTM technique has begun to be used in error forecasting models. Alizadeh et al. (2021) developed a Self-Activated Internal Attention based LSTM model and used it for four catchments with various hydroclimatic conditions across the United States. The attention based LSTM is designed to account for the dynamics of time series by focusing on the most informative and ignoring the less relevant time steps of a sequence. The main objective of this work is to develop and test LSTM-based error forecasting models for output assimilation of the conceptual model GRD (Jay-Allemand et al., 2019) across a range of lead times (up to 12 [H]). Runoff errors associated with input data are not treated in this study since it is considered that at the short lead times of this study hydrological uncertainties (i.e. uncertainties related to R-R models) are dominant (Demargne et al., 2010). The number of study catchments is very limited in this work. Only three catchments with an identical hydrological regime (Mediterranean) are investigated. It is however tried to answer the following questions:

1. How do error forecasting performances evolve with increasing lead time?
2. How single step forecasting is evaluated against multiple step forecasting?
3. Does training the model only based on errors of high flows improve error forecasting?
4. Does error forecasting using more explanatory variables bring performance improvement?

2 Case study

2.1 Catchments and data

The case study is based on two gauged catchments of the Loup River and one gauged catchment of the Brague River located in the West of the Alpes-Maritimes department in the South of France. These catchments belong to the Mediterranean regime and are shown in Figure 1. Their key information is listed in Table 1 and their average monthly and interannual observed discharge and rainfall are shown in Figure 2. Observed time series of discharge for the Loup and the Brague rivers are available from the Banque Hydro date set (http://hydro.eaufrance.fr/) at variable time steps over a 15-year period from 2006 to 2020. These data are subsequently interpolated to the common time step 15 [min] by Organde (2021). Rainfall observations consist of two consecutive segments from different sources. The first is provided by Novimet (https://www.novimet.com/) and includes observations collected since 2015 by the Hydrix and Collobrière radar products of Météo-France at time steps 1 [hour] and 15 [min]. The

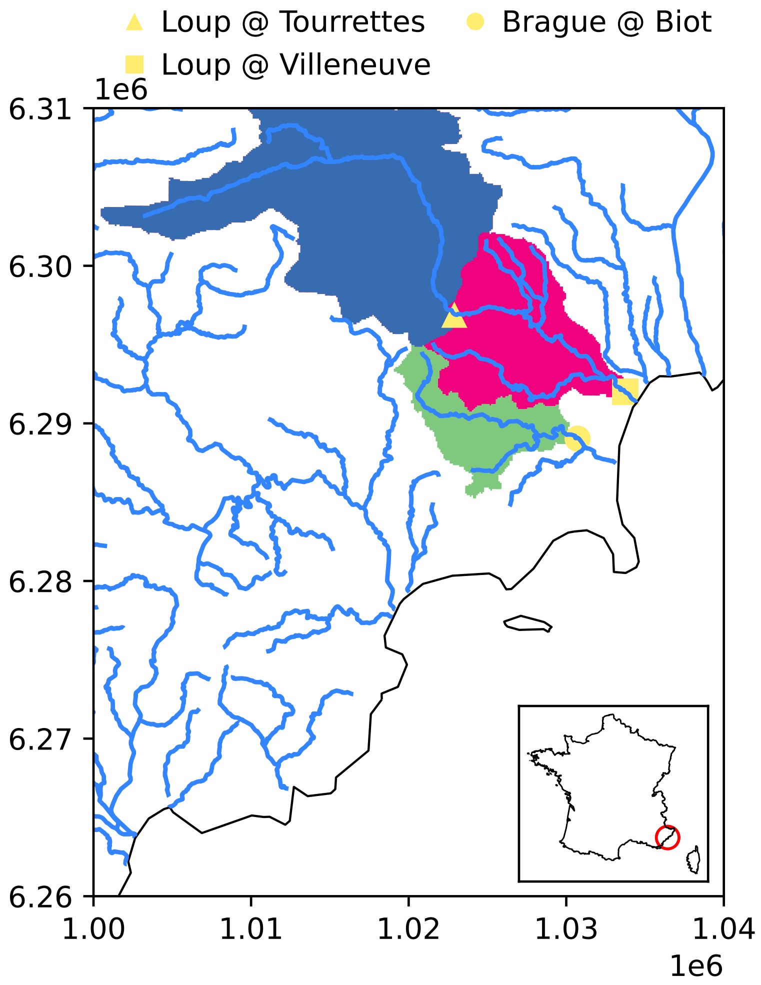

Figure 1: Location of the 3 case study catchments.

Catchment	Area [km^2]	Missing-flow days [%]	Missing-rain days [%]	0-flow days [%]	Mean flow [$\frac{\mathrm{mm}}{\mathrm{year}}$]	Runoff ratio [-]
Brague	41	15.29	0.73	18	338	0.35
Tourrettes	206	1.65	0.73	2	658	0.58
Villeneuve	289	0.09	0.73	0	538	0.49

Table 1: Case study catchments and their properties.

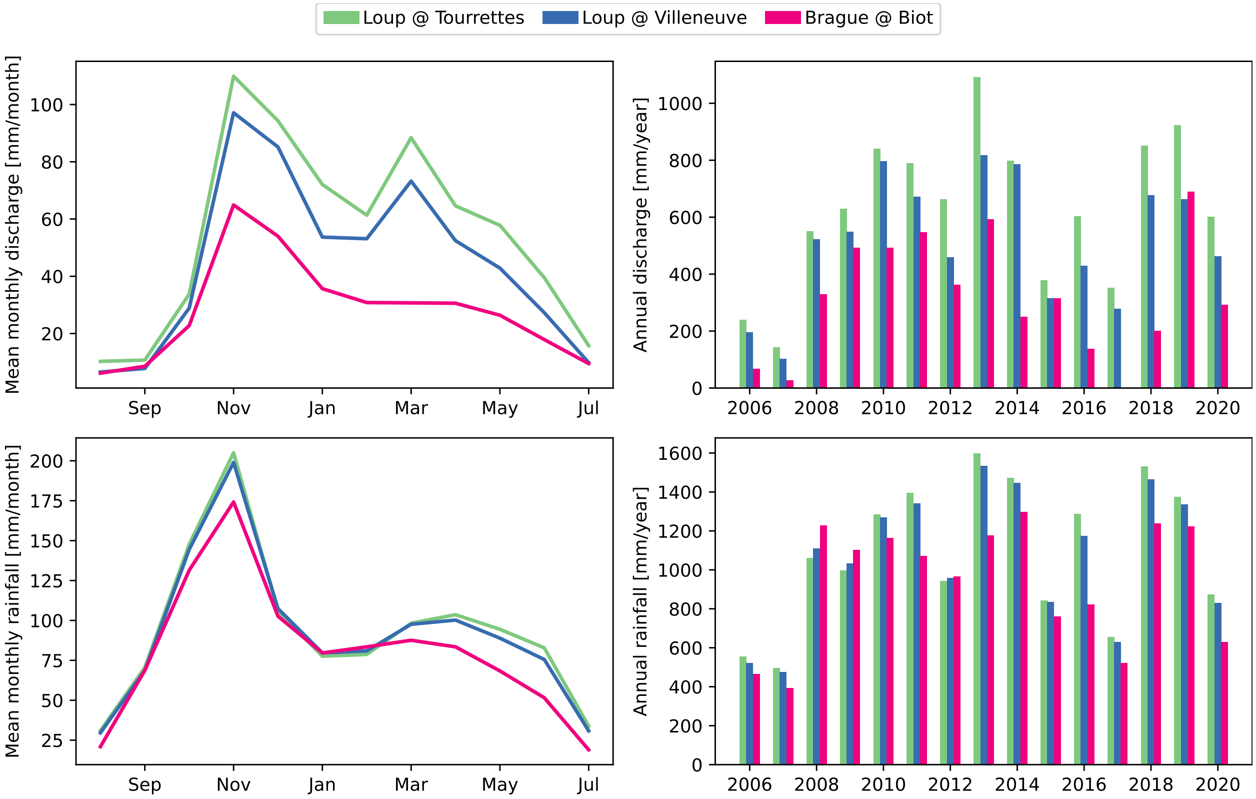

Figure 2: Mean observed streamflow and rain for each month of the year (panels on the left) and for individual years (panels on the right). In right panels years with missing data are taken into account while in left panels only full-record years are considered.

second contains rainfall observations from the Météo-France Antilope J+1 product at the hourly and 15 [min] time steps. These observations are corrected and reanalyzed with respect to gauge measurements. The two segments are concatenated to get continuous rainfall data over the period 2006 to 2020 (Organde, 2021).

2.2 Conceptual rainfall-runoff model

Organde (2021) used the GRD conceptual distributed R-R model (Jay-Allemand et al., 2019) to simulate sub-hourly (15 [min]) streamflows in the case study catchments for the period 2006 to 2020. The implemented GRD variant represents interception, infiltration, and percolation and has four calibration parameters. For further details regarding the modeling process, please refer to Organde (2021). The present study aims to assimilate these simulated discharges.

3 METHODS

At time step t, the simulation error (e^t) and the corrected simulated discharge (Q^t_{asim}) are defined as follows:

$$e^t = Q^t_{\mathrm{obs}} - Q^t_{\mathrm{grd}} \tag{1}$$

$$Q^t_{\mathrm{asim}} = e^t + Q^t_{\mathrm{grd}} \tag{2}$$

The idea here is to use an LSTM-based model to predict GRD's output error H time steps [hour] into future.

LSTM networks are a type of Recurrent Neural Networks (RNNs) and are able to learn time dependency in time series prediction problems. This an appealing feature that traditional neural networks like ANNs do not have and domains such as hydrological forecasting can benefit from it. Please refer to Hashemi et al. (2021) for an overview of the LSTM's principles.

3.1 Model training

The goal is to forecast time-varying simulation error ($[\mathrm{m}^3/\mathrm{s}]$) as target, $y = (y^1, y^2, ..., y^t, ..., y^T) \in \mathbb{R}^T$, H time steps into future given one or several (M) explanatory variables as features, $X = (X_1, X_2, ..., X_M)$. In mathematical terms, the task consists of learning one optimal set of parameters, θ, of the model Φ, so that the targets (forecasted simulation errors $[\mathrm{m}^3/\mathrm{s}]$), $\hat{y}$, accurately approaches the real targets (y) and thus the loss function $l(\hat{y}, y)$ is globally minimized:

$$\hat{y} = \Phi_\theta(X^{t-L+1}, ..., X^{t-1}, X^t, \theta) \tag{3}$$

Where L [hour] is the LSTM's window size for looking to the past and is hereafter called lookback. All LSTM-based models of this study are trained using the mean square error (*mse*) as loss function.

Model training is performed two times. Once taking into account the whole flow range and once on only high flows (i.e. flows lying in the 3rd and 4th quartiles). The two trainings are differentiated hereafter by the terms "thresholding is False" (in the former case) and "thresholding is True" (in the latter case).

Typically, three sets of data, namely, training, validation, and test sets, are needed to perform a DL task. The training data is used to fit the model, i.e. finding model optimal weights and biases. The validation set is used to provide an unbiased evaluation of the model performance during the learning process and to prevent over-fitting. The test set is used to provide an unbiased evaluation of the final model fit based on unseen data. Deciding the ratios between training, validation, and test sets is very dependent on the problem and data available. In this study, the common ratios 60%, 20%, and 20% are used to get the train (period P1), validation (period P2), and test (period P3) sets, respectively.

3.2 Choice of hyperparameters

In a DL task hyperparameters control the learning process and need therefore to be tuned. Hyperparameters are many. To avoid taking some a priori values of the hyperparameters and at the same time to keep the tuning task within manageable proportions, the hyperparameter optimization is conducted only for the following three hyperparameters, batch size, lookback, and lead time. The search parameter space is given therefore based on the following values for the chosen hyperparameters:

- lead times (H): 1, 3, 6, 12 [hour]
- lookbacks (L): 1, 3, 6, 12 [hour]

- batch sizes: 128, 256, 512

Every possible combination of these values are tested. The combination that achieves the highest performance on test data (period P3) is considered as the optimal hyperparameter set.

Taking into account the geometry hyperparameters, the number of LSTM layers and nodes are set to 1 and 32, respectively. The number of epochs and patience depend on the model variant and will be specified later when describing the two tested LSTM variants. Taking into account the regime of the catchments, study lead times for each catchment are selected based on autocorrelation coefficients between e^t and e^{t+H} (Table 2). Lead times associated with too small (< 0.5) autocorrelation coefficients are not studied.

Horizon [hour]	Brague at Biot	Loup at Tourrettes	Loup at Villeneuve
1	0.97	0.99	0.99
3	0.80	0.94	0.95
6	0.55	0.83	0.86
12	0.36	0.62	0.71
24	0.22	0.31	0.45
48	0.13	0.15	0.27

Table 2: The Pearson correlation coefficients between e^t and e^{t+H} for different lead times and for different catchments.

3.3 Model variants

3.3.1 With respect to the configuration of features

Three model variants with respect to inputs are considered. The first variant ("UV") is univariate and uses only discharge error (e^t) as input. In the second variant ("MV1"), which is a multivariate model, discharge observations (Q^t_{obs}) are added as additional input, aiming to improve error forecasting accuracy. The third variant ("MV2") includes all input variables from the second variant (e^t and Q^t_{obs}) plus rainfall observations (P^t_{obs}) in order to investigate whether introducing this information brings further prediction accuracy improvement.

The data is standardized using the mean and standard deviation of only the training period (P1). This is because the model should have no access to the values in the validation and test sets to prevent data leakage.

3.3.2 With respect to the LSTM's topology

Two variants with respect to the topology of the LSTM are tested. First, the standard many-to-one topology (Figure 3) and then going forward with the more complex many-to-many (or sequence-to-sequence) structure, also called **E**ncoder-**D**ecoder LSTM (**ED**-LSTM).

The standard LSTM takes a sequence of past information $(X^{t-L+1}, ..., x^{t-1}, x^t)$ and predicts the target lead time (H, for example, H=6 [hour]) into future (y^{t+H}). The forecast (lead time) window has a width of 1 and is located at time step $t + H$. It is therefore a single step ("SS") prediction since the prediction is made for only one time step. The length of the sequence of the past data is the lookback size (L). The sequence-to-sequence LSTM makes prediction for all consecutive time steps of a future sequence all at once (from y^{t+1} to y^{t+H}). It is therefore a multi step ("MS") prediction. The forecast (lead time) window has a width of H. We call hereafter the width of this window the forecast "horizon" defining the span of

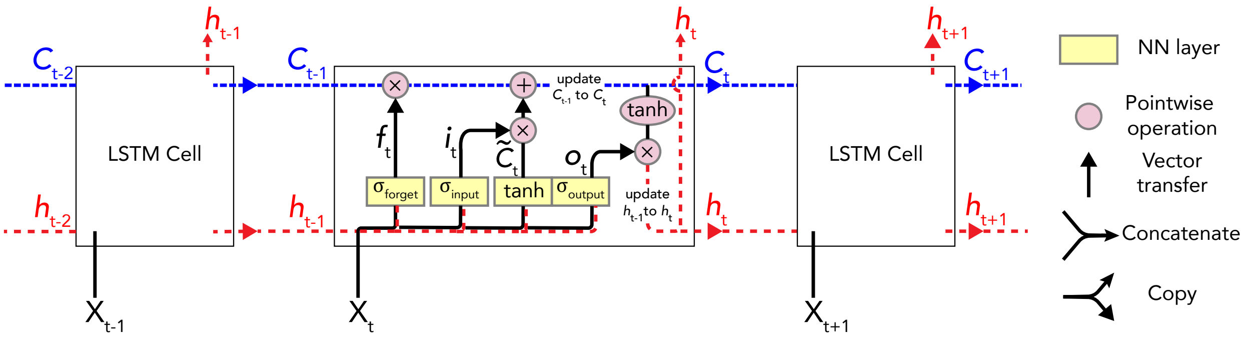

Figure 3: Time-unfolded building blocks of a standard LSTM network.

future prediction. The multi step forecaste has the practical advantage of making predictions in a single shot for lead time values from 1 to H ([hour]). In an ED-LSTM, the encoder part first converts the inputs in the lookback window $(x^{t-LB+1}, ..., x^{t-1}, x^{t})$ to a fixed length vector (i.e. context vector), serving as a summary of the information in the past. The context vector as well as the last encoder state are then fed to the decoder part to predict the value of the steps in the lead time window. The architecture of this model thus involves a first LSTM layer (in the encoder part), a repeat vector layer (for connecting the encoder and decoder part), a second LSTM layer (in the decoder part), and a Dense Time Distributed layer (in the decoder part) to separate, for each time step, the mixed output received from the decoder. Table 3 presents a summary of the different investigated combinations of variants (with respect to both inputs and topology of LSTM) as well as the maximum number of epochs ($\mathrm{N}_{\mathrm{e,max}}$) and patience used for each of them. For MS variants a greater patience and $\mathrm{N}_{\mathrm{e,max}}$ is taken into account since the problem becomes more complex.

Model	Feature structure	Features	Output structure	Output(s)	Patience	$\mathrm{N}_{\mathrm{e,max}}$
SS-UV	Univariate	e^t	A single time step	y^{t+H}	10	200
SS-MV1	Multivariate	e^t, Q^t_{obs}	A single time step	y^{t+H}	10	200
SS-MV2	Multivariate	e^t, Q^t_{obs}, P^t_{obs}	A single time step	y^{t+H}	10	200
MS-UV	Univariate	e^t	Multiple time steps	$(y^{t+1}, y^{t+2}, ..., y^{t+H})$	15	250
MS-MV1	Multivariate	e^t, Q^t_{obs}	Multiple time steps	$(y^{t+1}, y^{t+2}, ..., y^{t+H})$	15	250
MS-MV2	Multivariate	e^t, Q^t_{obs}, P^t_{obs}	Multiple time steps	$(y^{t+1}, y^{t+2}, ..., y^{t+H})$	15	250

Table 3: Summary of the implemented error forecast LSTM-based models.

3.4 Model benchmarking

Performances of the LSTM-based models are compared against a “Naive” model that assumes that the H-step ahead error (e^{t+H}) equals the last observed error (e^t). Although being very simple, since errors of conceptual R-R models are generally strongly autocorrelated this benchmark model gives usually good results (Anctil et al., 2003).

3.5 Performance evaluation

The evaluation of the performance of the models is based on the root mean square error (RMSE) metric on the test data set (period P3).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{T}(Q^i_{\mathrm{sim}} - Q^i_{\mathrm{obs}})^2}{T}} \tag{4}$$

Where index $_\text{sim}$ in Q^i_sim refers to one of the three types of simulated discharges: 1) initial discharges of GRD (in short “Initial”), or 2) discharges of GRD corrected using an LSTM-based model, or 3) discharges of GRD corrected using the Naive model.

4 RESULTS AND DISCUSSION

The RMSE metric obtained from the optimal hyperparameter set of the LSTM-based, Naive, and Initial models is reported for each horizon in Table 4 (for when thresholding is False) and Table 5 (for when thresholding is True). The two tables present the results corresponding to exactly the same time steps. The RMSE results of all other hyperparameter sets for the LSTM-based, Naive, and Initial models are provided in Appendices A, B, and C, respectively. Note that

1. performances of the Naive and Initial models change with changing lookbacks since different lookbacks lead to different sets of time steps. In a similar way and due to the presence of NANs, time steps for the LSTM-based models vary depending on whether the model is an SS or an MS variant. For the sake of comparison, the considered time steps for the Naive and Initial models change correspondingly.
2. performances of the Naive and Initial models do not depend on the configuration of features.
3. all presented results correspond to the time steps in common between the two training approaches (when thresholding is True and when thresholding is False).

Catchment	Horizon [H]	UV		MV1		MV2		Naive		Initial	
		SS	MS	SS	MS	SS	MS	SS	MS	SS	MS
Biot	1	1.67	1.43	1.30	1.37	1.31	1.52	1.44	1.46	2.89	2.94
	3	2.40	2.24	2.40	2.31	2.15	2.08	2.50	2.53	2.85	2.88
	6	2.69	2.67	2.62	2.67	2.50	2.59	3.35	3.37	2.82	2.84
Tourrettes	1	1.19	1.50	1.20	1.54	1.14	1.55	1.22	1.27	8.56	8.92
	3	2.28	2.41	2.26	2.46	2.14	2.50	2.47	2.57	8.53	8.87
	6	3.46	3.55	3.48	3.39	3.34	3.48	3.85	3.95	8.50	8.79
	12	4.76	4.74	4.75	4.66	4.67	4.67	5.59	5.60	8.61	8.65
Villeneuve	1	0.85	1.16	0.99	1.52	1.05	1.49	1.10	1.10	6.74	6.79
	3	2.51	2.45	2.65	2.43	2.61	2.67	2.74	2.76	6.65	6.66
	6	3.75	3.82	3.75	3.49	3.87	3.54	4.23	4.26	6.59	6.58
	12	4.58	4.75	4.49	4.49	4.66	4.63	5.43	5.46	6.54	6.56

Table 4: Comparison of the minimum RMSE obtained from different LSTM-based models as well as the Naive and Initial GRD for different horizons and when **thresholding is False**.

Catchment	Horizon [H]	UV		MV1		MV2		Naive		Initial	
		SS	MS	SS	MS	SS	MS	SS	MS	SS	MS
Biot	1	1.70	1.72	1.40	1.46	1.35	1.44	1.44	1.46	2.89	2.94
	3	2.33	2.33	2.35	2.28	2.22	2.12	2.50	2.53	2.85	2.88
	6	2.64	2.75	2.60	2.67	2.55	2.57	3.35	3.37	2.82	2.84
Tourrettes	1	1.20	1.49	1.22	1.58	1.09	1.64	1.22	1.27	8.56	8.92
	3	2.27	2.49	2.29	2.52	2.21	2.49	2.47	2.57	8.53	8.87
	6	3.53	3.47	3.49	3.58	3.50	3.46	3.85	3.95	8.50	8.79
	12	4.81	4.71	4.67	4.77	5.06	4.84	5.59	5.60	8.61	8.65
Villeneuve	1	0.87	1.35	1.09	1.38	1.18	1.46	1.10	1.10	6.74	6.79
	3	2.39	2.40	2.70	2.39	2.90	2.31	2.74	2.76	6.65	6.66
	6	3.84	3.79	3.69	3.70	3.93	3.50	4.23	4.26	6.59	6.58
	12	4.62	4.79	4.50	4.83	5.15	5.18	5.43	5.46	6.54	6.56

Table 5: Comparison of the minimum RMSE obtained from different LSTM-based models as well as the Naive and Initial GRD for different horizons and when **thresholding is True**.

Based on the presented results the following observations can be made.

1. No matter the catchment and the horizon and whether thresholding is applied or not, both LSTM-based models and Naive models do always manage to improve initial simulations of the GRD model. These improvements are more pronounced for the Loup at Tourrettes catchment and less significant for the Brague at Biot catchment.

2. Comparing the LSTM-based models with the Naive model, there is always an LSTM-based model outperforming the Naive model. This is true for the three catchments and both training approaches. More precisely, for horizons as small as 1 [hour], the Naive model gives performances very close to those given by the LSTM. This is natural taking into account that catchments are from the Mediterranean regime and the Naive model has an autocorrelation nature. With increasing horizon (H$\geq$ 3 [hour]), the LSTM-based models start to clearly out perform, almost always, independent of their input configuration and their LSTM topology.

3. Comparing performances of different LSTM-based models, it is observed that they depend explicitly in the first place on horizon. Therefore, for each catchment, an LSTM model involving whatever configuration of inputs or whatever LSTM topology has always a better forecast performance at horizon H_1 than another LSTM model with a more/less complex configuration of inputs/LSTM topology at horizon H_2 that is greater than H_1. This behavior can be explained again by the fact that these are Mediterranean catchments featured by very short term responses.

4. Keeping the topology variant (SS or MS) constant and considering different input variants (UV, MV1, and MV2), all catchments do not show the same behavior. For the Brague at Biot catchment, with increasing horizon the most complex variant (MV2) outperforms for both SS and MS topologies. For the Loup at Tourrettes catchment and the SS models, the MV2 variant is better able to forecast with increasing horizon. This is while in MS models the simpler input variant UV (and MV1) gives better forecasts. The observed pattern is completely inverse in the Loup at Villeneuve catchment: the univariate variant UV in the SS models and the multivariate MV2 in the MS models outperform with increasing horizon. One possible explanation for such difference is the size of train data available for each catchment. Looking at Table 1, the Brague at Biot has the smallest amount of data. That may be why in this catchment feeding the LSTM with more features improves forecasts. The rate of missing discharges in the Loup at Tourrettes and the Loup at Villeneuve catchments supports even more this speculation: the latter having a very small missing data rate favors variants with less inputs. The former (Tourrettes) that has a rate of missing discharges smaller than Biot but larger than Villeneuve turns out to behave in part in accordance with Brague (more features in the SS models) and in part consistent with Villeneuve (less features in the MS models).

5. Between SS and MS variants, the latter have a harder task to deal with. It is not however always the case that the SS models outperform the corresponding MS models. Outperformance of the SS models turns out also to depend on the training approach (whether thresholding on high flows or not). For instance, in the Brague at Biot catchment, when thresholding is False and the input configuration is univariate, the MS model outperforms at all horizons. But when thresholding is true, it is the SS model that outperforms at all horizons. Taken together, the SS models outperform the corresponding MS models at either most (all) of the horizons or at half of the studied horizons. Nevertheless, the difference in RMSE between two corresponding SS and MS models is not so important in any case that the MS model and the advantage it offers should be excluded.

6. Comparing the RMSE between two exactly corresponding cases of Tables 4 (thresholding is False) and 5 (thresholding is True), no obvious conclusion can be made except that in 41 out of 66 cases

thresholding does not improve the RMSE (on exactly the same time steps).

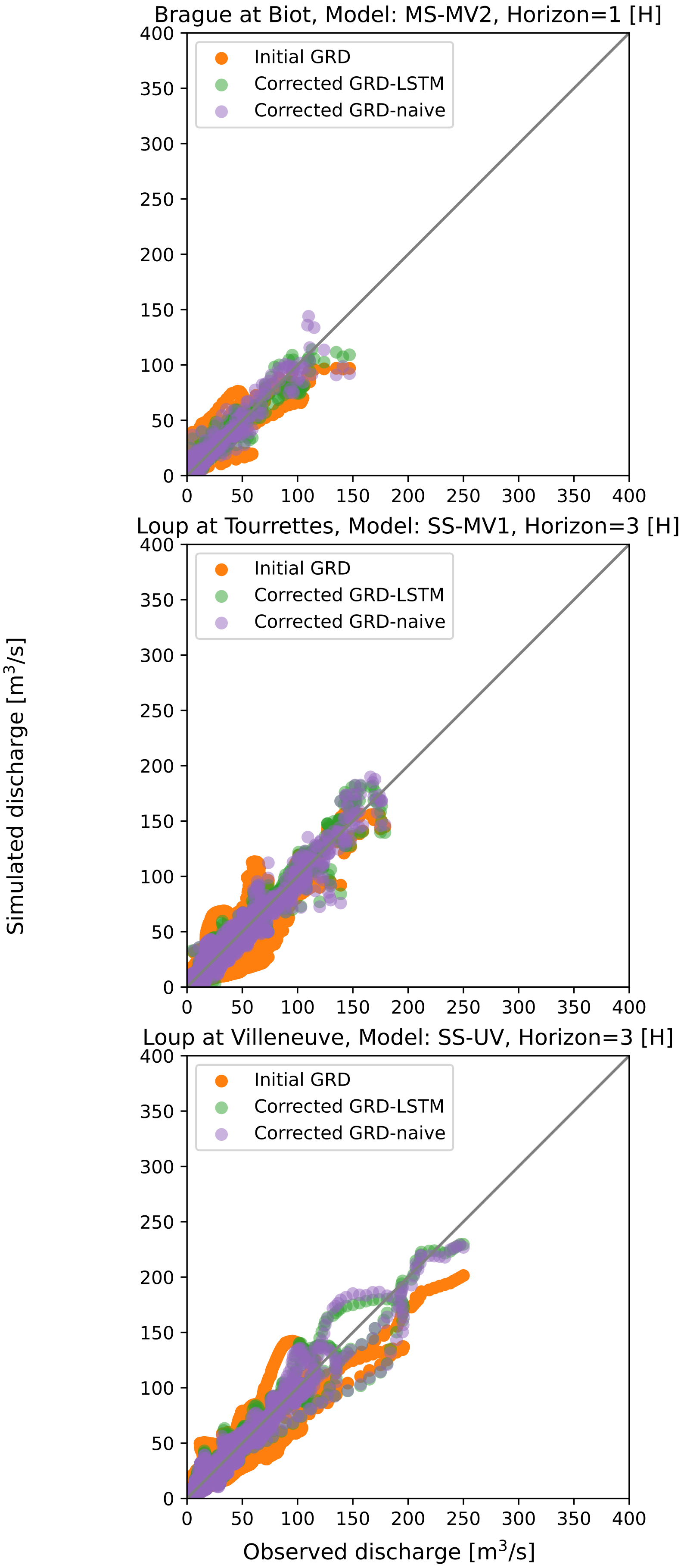

Figure 4: Scatter plot example of corrected discharges by the LSTM-based models versus the Naive model versus the Initial simulations by the GRD model.

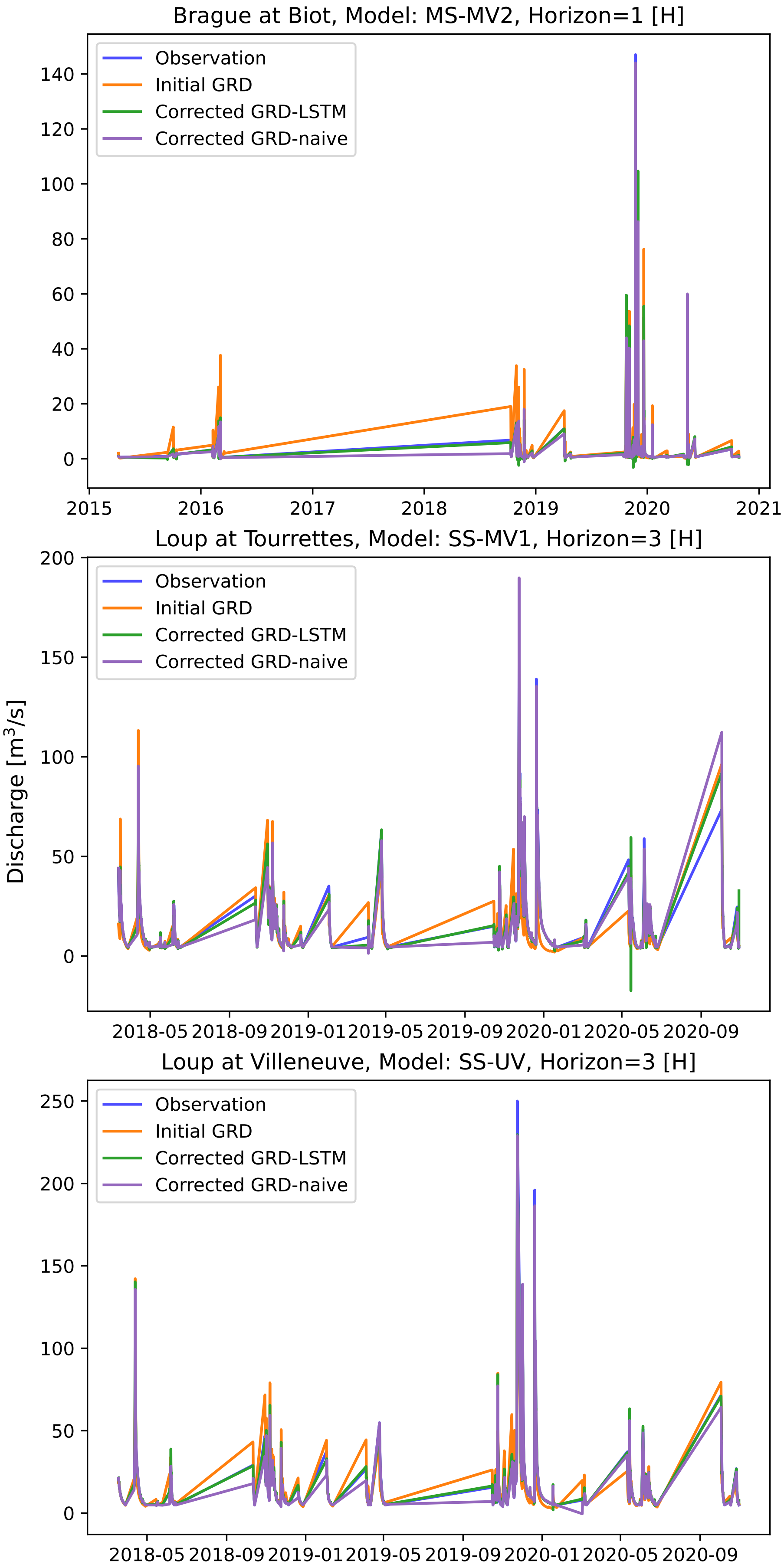

Figure 5: Hydrograph example of corrected discharges by the LSTM-based models versus the Naive model versus the Initial simulations by the GRD model.

5 CONCLUSION

In this study, discharge simulations for three Mediterranean catchments were corrected using LSTM-based models. The discharge simulations had been previously carried out at the time step 15 $[\mathrm{min}]$ using the GRD model: a conceptual fully distributed rainfall-runoff model. The discharge correction approach consisted of forecasting simulation error of the GRD model H time steps into future. The LSTM models varied in two aspects: 1) the topology of the LSTM, 2) the choice of explanatory variables (features). A range of lead times was investigated: from 1 to 12 [hour]. The two tested topologies included the standard LSTM model yielding forecaste of a single time step at a time and the Encode-Decoder LSTM yielding a sequence of forecast all at once. The LSTM-based models were first trained on the whole range of available discharges and, in a second time, on only flows lying in the 3rd and 4th quartiles of discharge. LSTM corrections obtained from different trainings, topologies, and features were compared against each other and a naive model assuming that the H-step ahead error equaled the last observed error. The results suggest the following conclusions:

1. The optimal lead time depends on the regime of catchment. In this study since all catchments belonged to the Mediterranean regime errors were highly autocorrelated and therefore with increasing lead time, forecast performances degraded monotonically. Therefore, the optimal lead time occurred always at the minimum lead time. In other regimes with a different response time and behavior, it is expected that the optimal lead time does not necessarily happen at the smallest lead time.

2. The size of available data and the complexity of features both tend to have complementary roles. That is, when there is ample data for model training less features give better forecasts. But when data is scarce, including more feature could make up for this issue.

3. Thresholding on high flows does not show an obvious improvement of the RMSE metric *for the time steps in common between the two cases (when thresholding is performed and not)*.

4. The multi step approach shows forecast performances (very) similar to single step forecast performances. The advantage of getting a sequential forecast for multiple lead times all at once should not be therefore excluded.

The current conclusions are drawn based on the results only from three catchments that is too few. It is recommended to validate them in a far larger sample containing catchments of different regimes. A future research perspective is to test the Attention based LSTM topology that makes the model free from the need to manually detect important dynamics of the time series (by for instance thresholding on high flows).

REFERENCES

M. Akbari and A. Afshar. Similarity-based error prediction approach for real-time inflow forecasting. *Hydrology Research*, 45(4-5):589–602, 11 2013.

B. Alizadeh, A. G. Bafti, H. Kamangir, Y. Zhang, D. B. Wright, and K. J. Franz. A novel attention-based lstm cell post-processor coupled with bayesian optimization for streamflow prediction. *Journal of Hydrology*, 601:126526, 2021.

F. Anctil, C. Perrin, and V. Andréassian. Ann output updating of lumped conceptual rainfall / runoff forecasting models 1. *JAWRA Journal of the American Water Resources Association*, 39(5):1269–1279, 2003.

V. Babovic, R. Caňizares, H. R. Jensen, and A. Klinting. Neural networks as routine for error updating of numerical models. *Journal of Hydraulic Engineering*, 127(3):181–193, 2001.

V. Bidwell and G. Griffiths. Adaptive flood forecasting: an application to the waimakariri river. *Journal of Hydrology (New Zealand)*, pages 1–15, 1994.

J. Demargne, J. Brown, Y. Liu, D.-J. Seo, L. Wu, Z. Toth, and Y. Zhu. Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmos. Sci. Lett.*, 11(2):114–122, Apr 2010.

M. Ghaith, A. Siam, Z. Li, and W. El-Dakhakhni. Hybrid hydrological data-driven approach for daily streamflow forecasting. *Journal of Hydrologic Engineering*, 25(2):04019063, 2020.

M. Goswami and K. M. O'Connor. A "monster" that made the smar conceptual model "right for the wrong reasons". *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 55(6):913–927, 2010.

R. Hashemi, P. Brigode, P.-A. Garambois, and P. Javelle. How can regime characteristics of catchments help in training of local and regional lstm-based runoff models? *Hydrology and Earth System Sciences Discussions*, 2021:1–33, 2021. doi: 10.5194/hess-2021-511. URL https://hess.copernicus.org/preprints/hess-2021-511/.

G. B. Humphrey, M. S. Gibbs, G. C. Dandy, and H. R. Maier. A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a bayesian artificial neural network. *Journal of Hydrology*, 540:623–640, 2016.

M. Jay-Allemand, P. Javelle, I. Gejadze, P. Arnaud, P.-O. Malaterre, J.-A. Fine, and D. Organde. On the potential of variational calibration for a fully distributed hydrological model: application on a mediterranean catchment. *Hydrology and Earth System Sciences*, 2019:1–24, 2019.

Y. Li, D. Ryu, A. W. Western, and Q. J. Wang. Assimilation of stream discharge for flood forecasting: Updating a semidistributed model with an integrated data assimilation scheme. *Water Resources Research*, 51(5):3238–3258, 2015.

P. Liu, X. Zhang, Y. Zhao, C. Deng, Z. Li, and M. Xiong. Improving efficiencies of flood forecasting during lead times: an operational method and its application in the baiyunshan reservoir. *Hydrology Research*, 50(2):709–724, 2019.

Y. Liu, Y. Ji, D. Liu, Q. Fu, T. Li, R. Hou, Q. Li, S. Cui, and M. Li. A new method for runoff prediction error correction based on ls-svm and a 4d copula joint distribution. *Journal of Hydrology*, 598:126223, 2021.

G. Y. Muluye. Improving long-range hydrological forecasts with extended kalman filters. *Hydrological sciences journal*, 56(7):1118–1128, 2011.

D. Organde. Étude pour le passage à une modélisation hydrologique continue au pas de temps infra-horaire (smash) intégrée à la plateforme rainpol®en vue de l'amélioration du système de prévision des crues du smiage maralpin. Technical report, HYDRIS hydrologie, 2021.

T. Pagano, Q. Wang, P. Hapuarachchi, and D. Robertson. A dual-pass error-correction technique for forecasting streamflow. *Journal of Hydrology*, 405(3-4):367–381, 2011.

P. Pokhrel, D. E. Robertson, and Q. J. Wang. A bayesian joint probability post-processor for reducing errors and quantifying uncertainty in monthly streamflow predictions. *Hydrology and Earth System Sciences*, 17(2):795–804, 2013.

S. Y. Schreider, A. Jakeman, B. Dyer, and R. Francis. A combined deterministic and self-adaptive stochastic algorithm for streamflow forecasting with application to catchments of the upper murray basin, australia. *Environmental Modelling & Software*, 12(1):93–104, 1997.

A. Y. Shamseldin and K. M. O'CONNOR. A real-time combination method for the outputs of different rainfall-runoff models. *Hydrological Sciences Journal*, 44(6):895–912, 1999.

L. Sun, O. Seidou, and I. Nistor. Data assimilation for streamflow forecasting: State-parameter assimilation versus output assimilation. *J. Hydrol. Eng.*, 22(3):04016060, 2017.

E. Toth, A. Montanari, and A. Brath. Real-time flood forecasting via combined use of conceptual and stochastic models. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 24 (7):793–798, 1999.

A The LSTM-based models

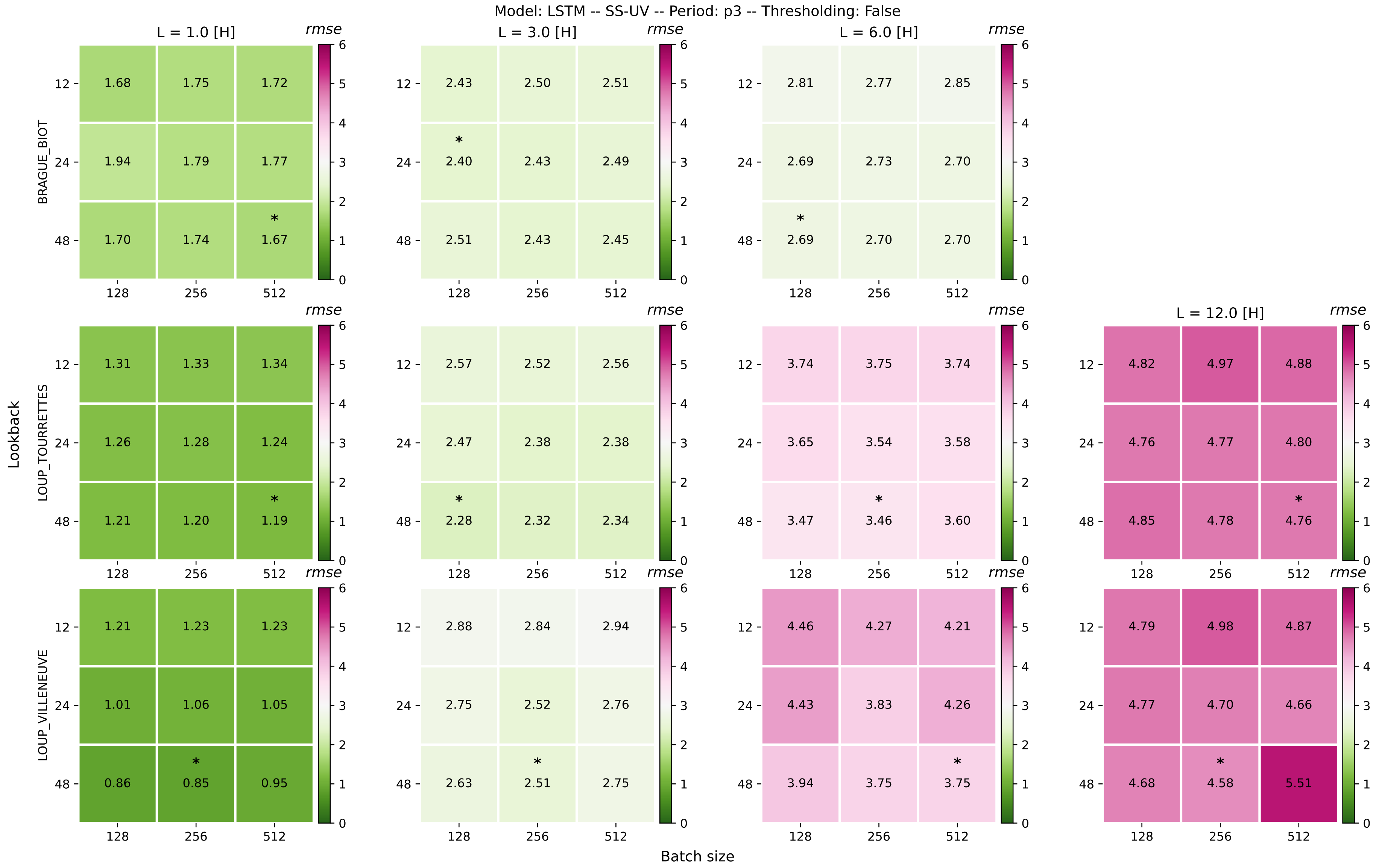

Figure 6: Results of the SS-UV model when thresholding is False.

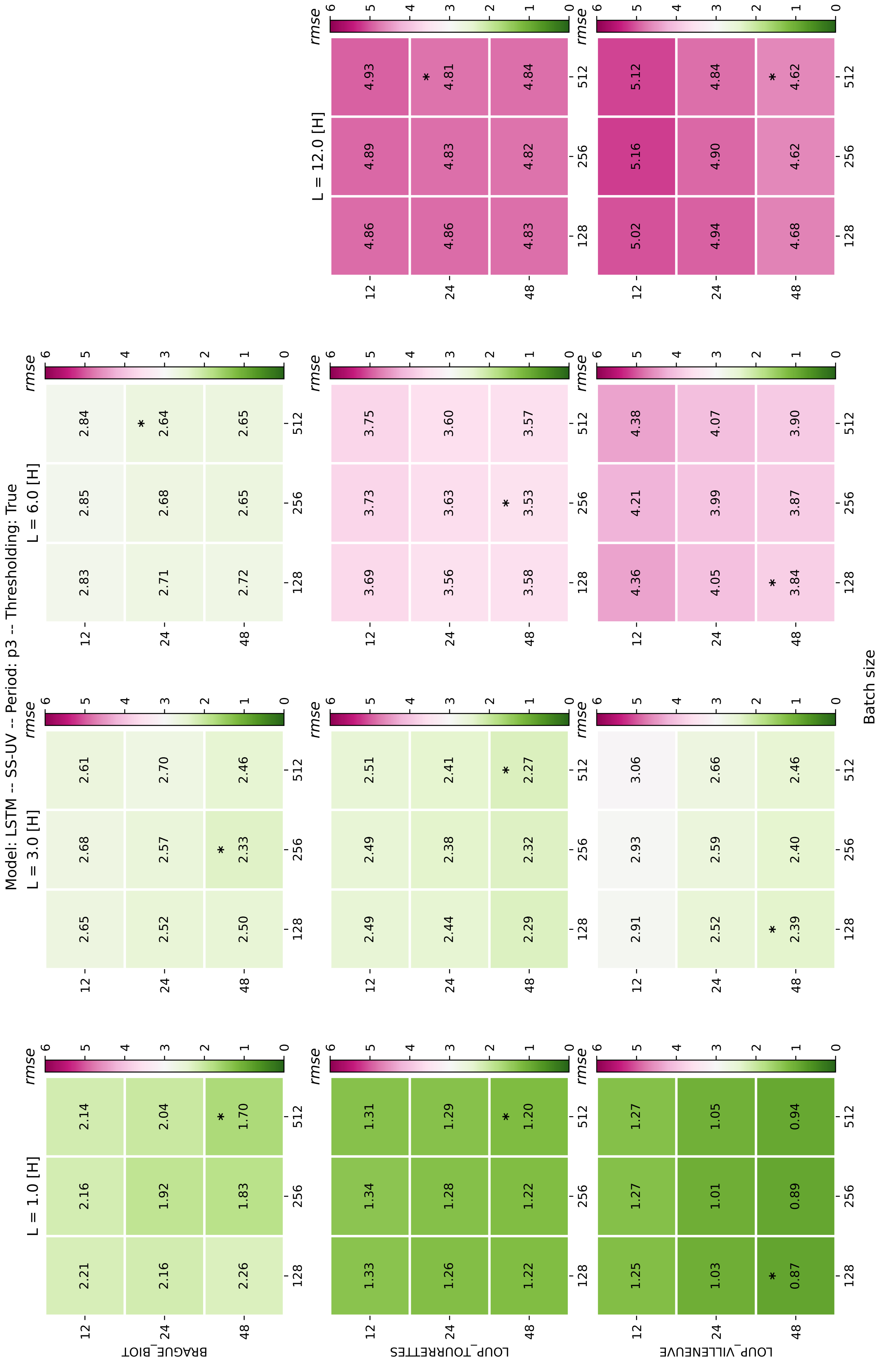

Figure 7: Results of the SS-UV model when thresholding is True.

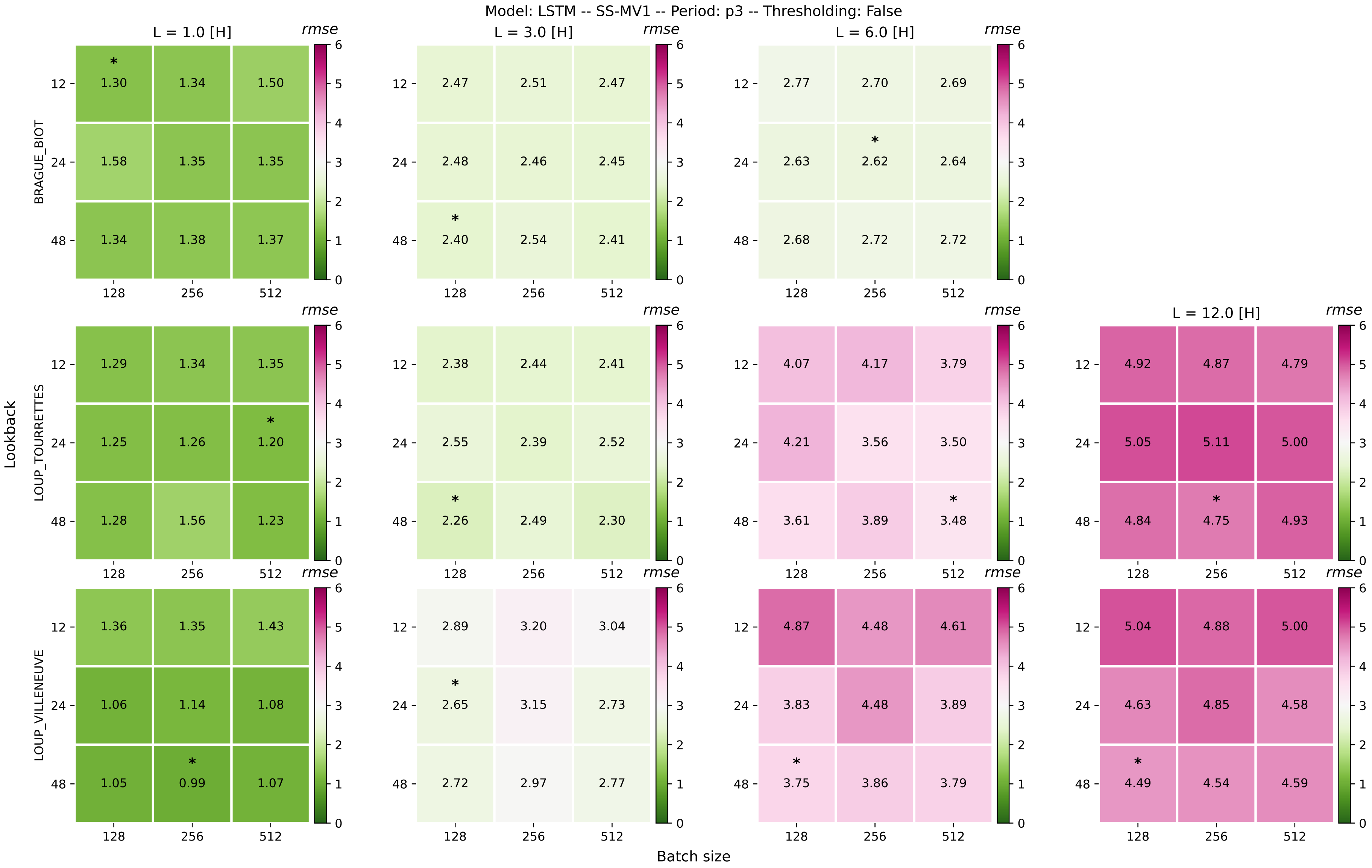

Figure 8: Results of the SS-MV1 model when thresholding is False.

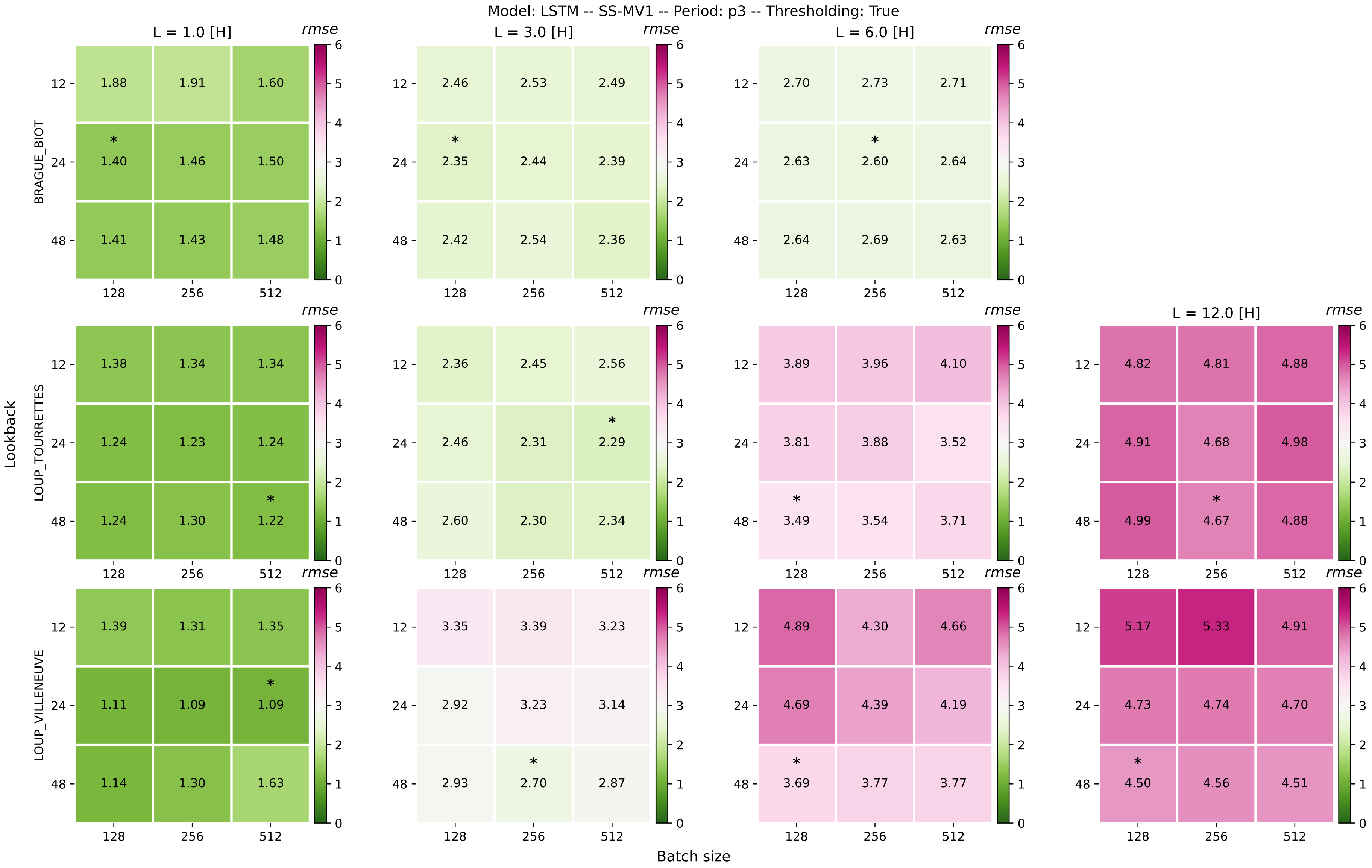

Figure 9: Results of the SS-MV1 model when thresholding is True.

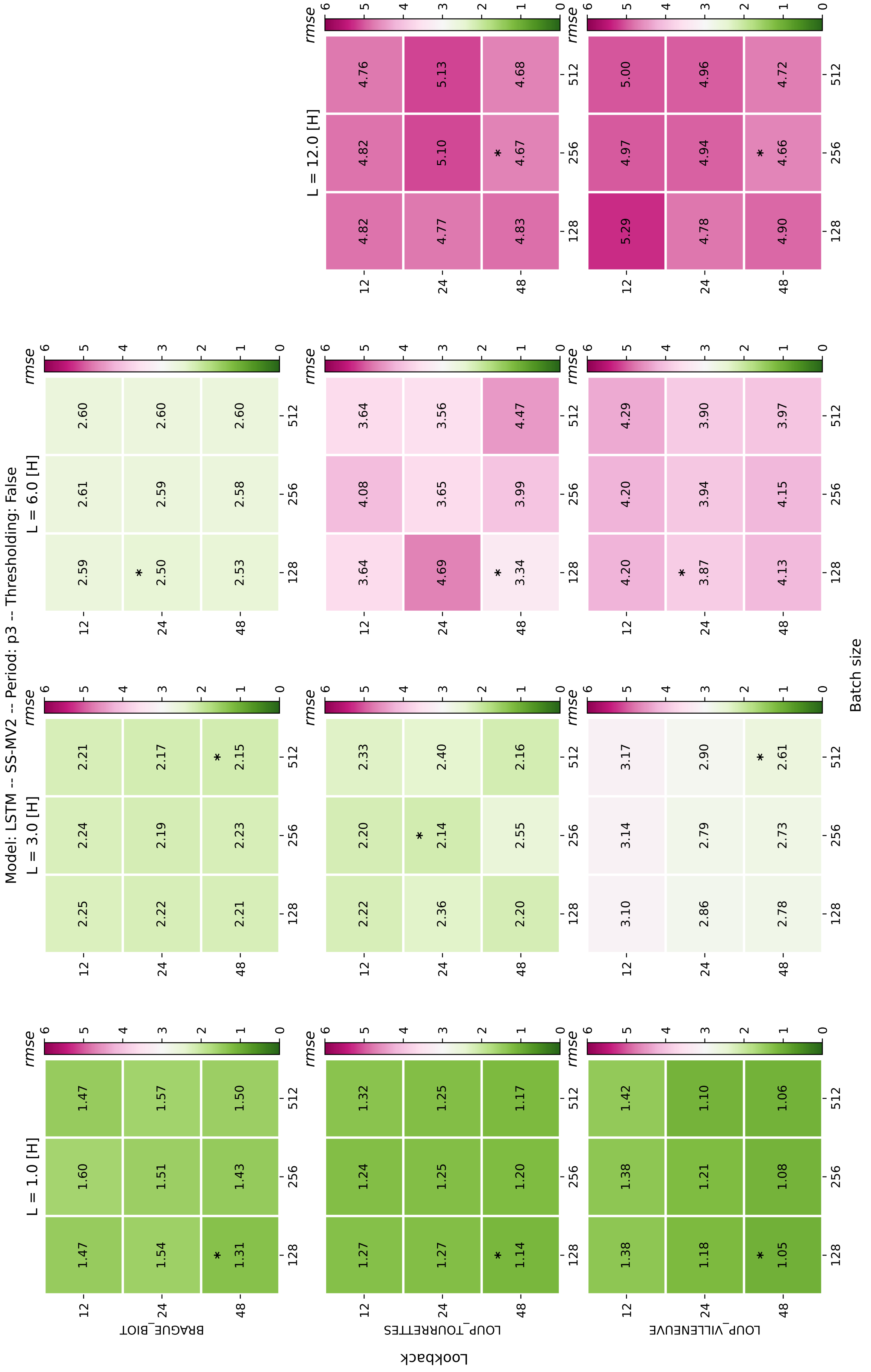

Figure 10: Results of the SS-MV2 model when thresholding is False.

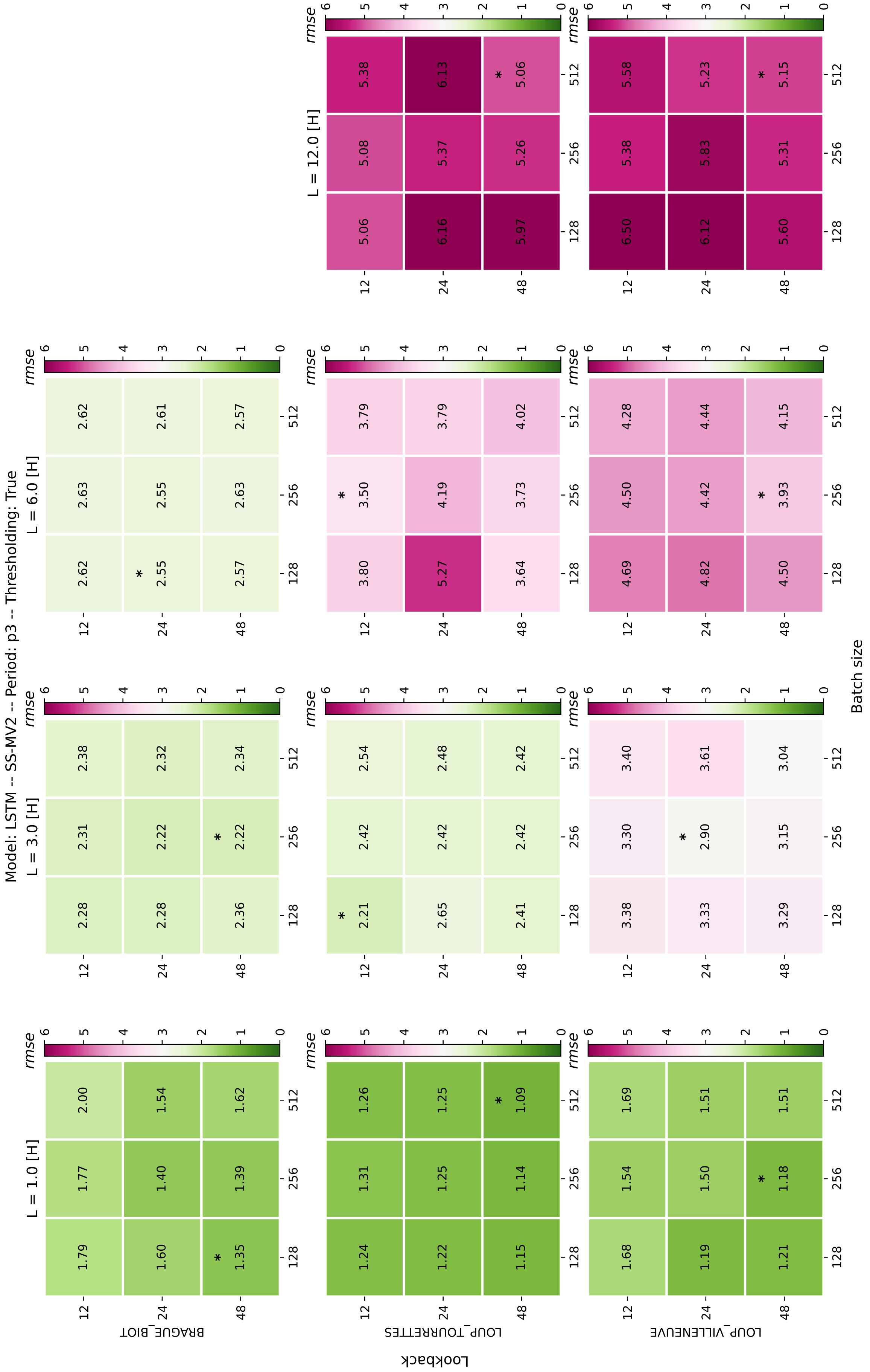

Figure 11: Results of the SS-MV2 model when thresholding is True.

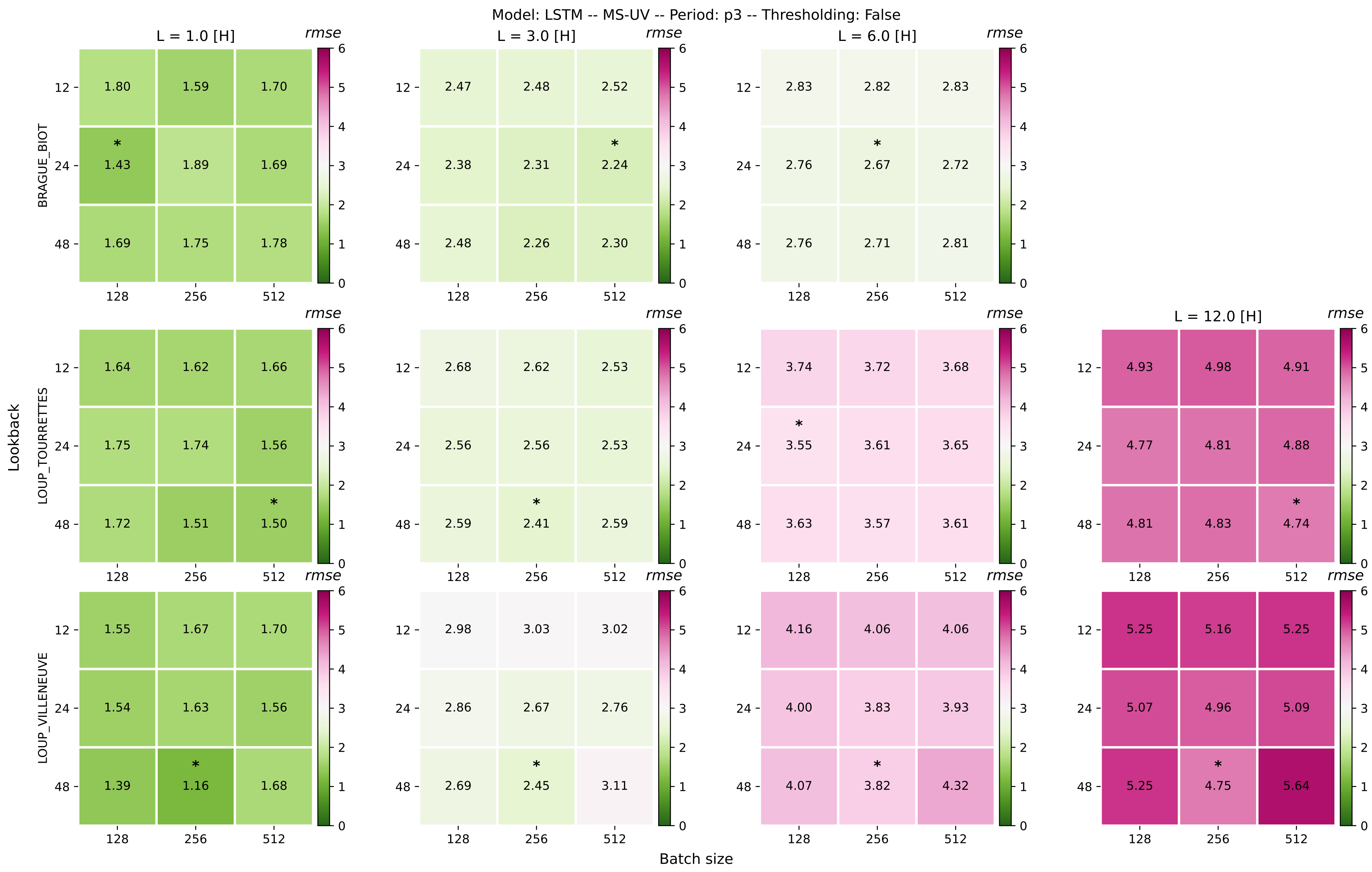

Figure 12: Results of the MS-UV model when thresholding is False.

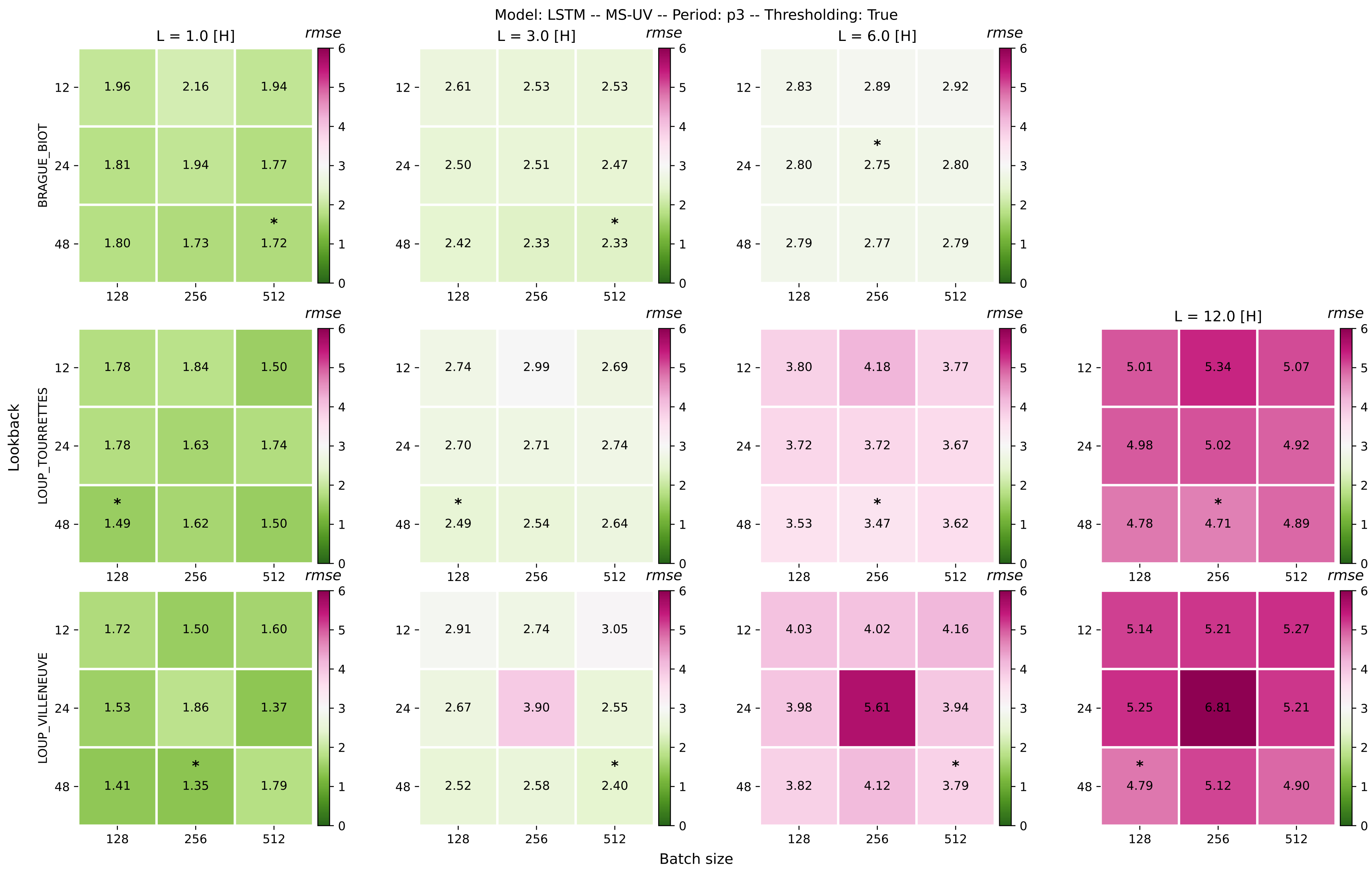

Figure 13: Results of the MS-UV model when thresholding is True.

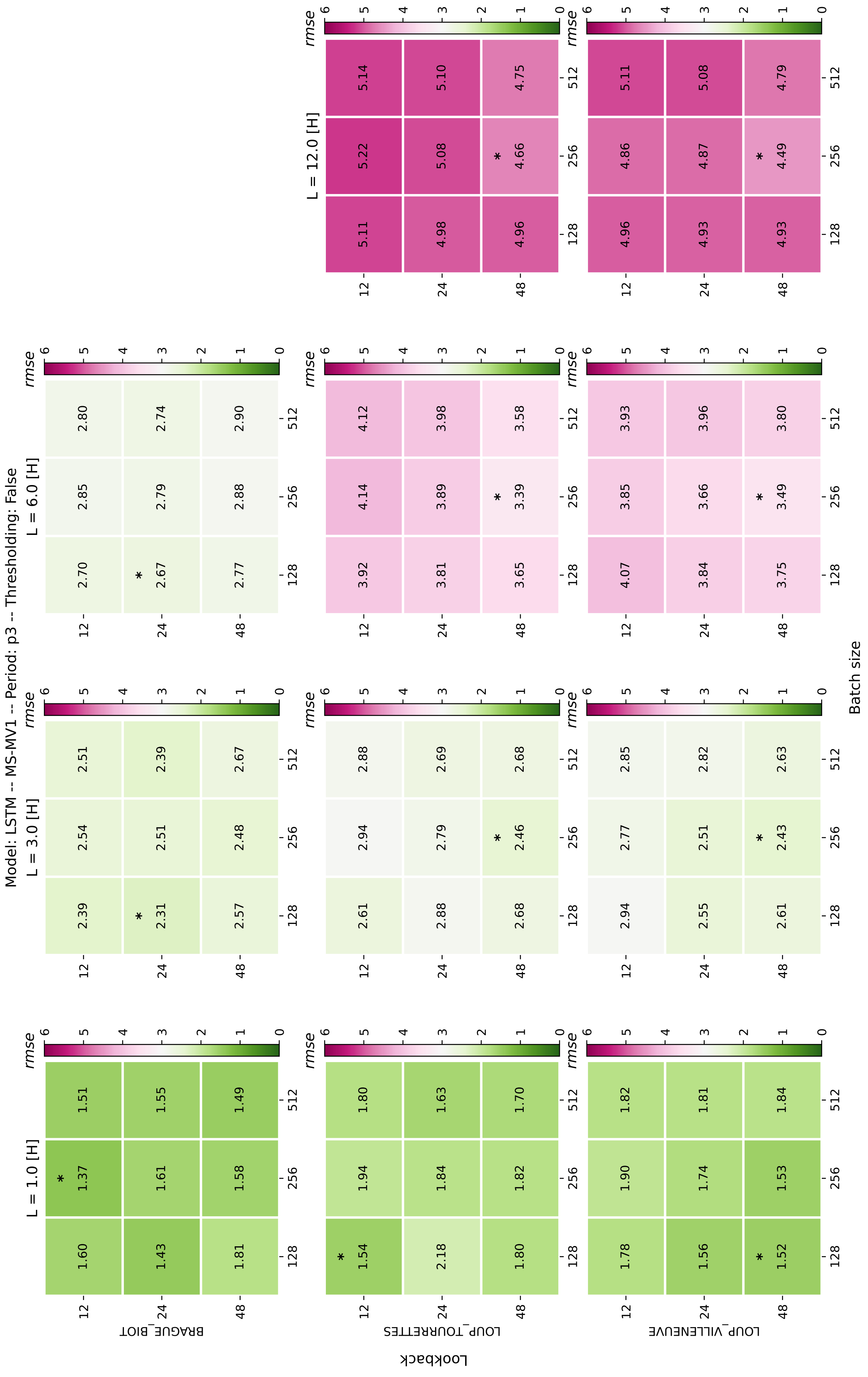

Figure 14: Results of the MS-MV1 model when thresholding is False.

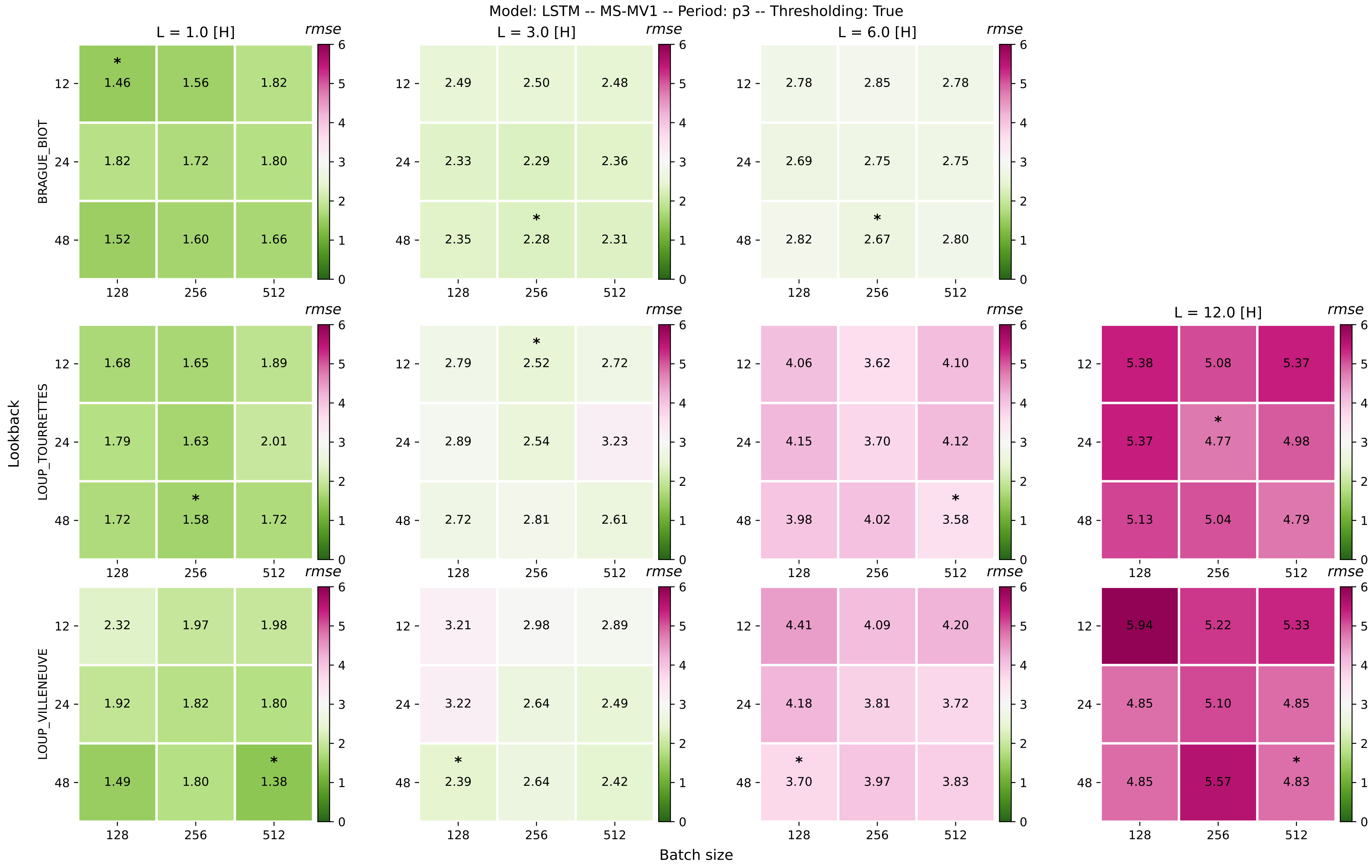

Figure 15: Results of the MS-MV1 model when thresholding is True.

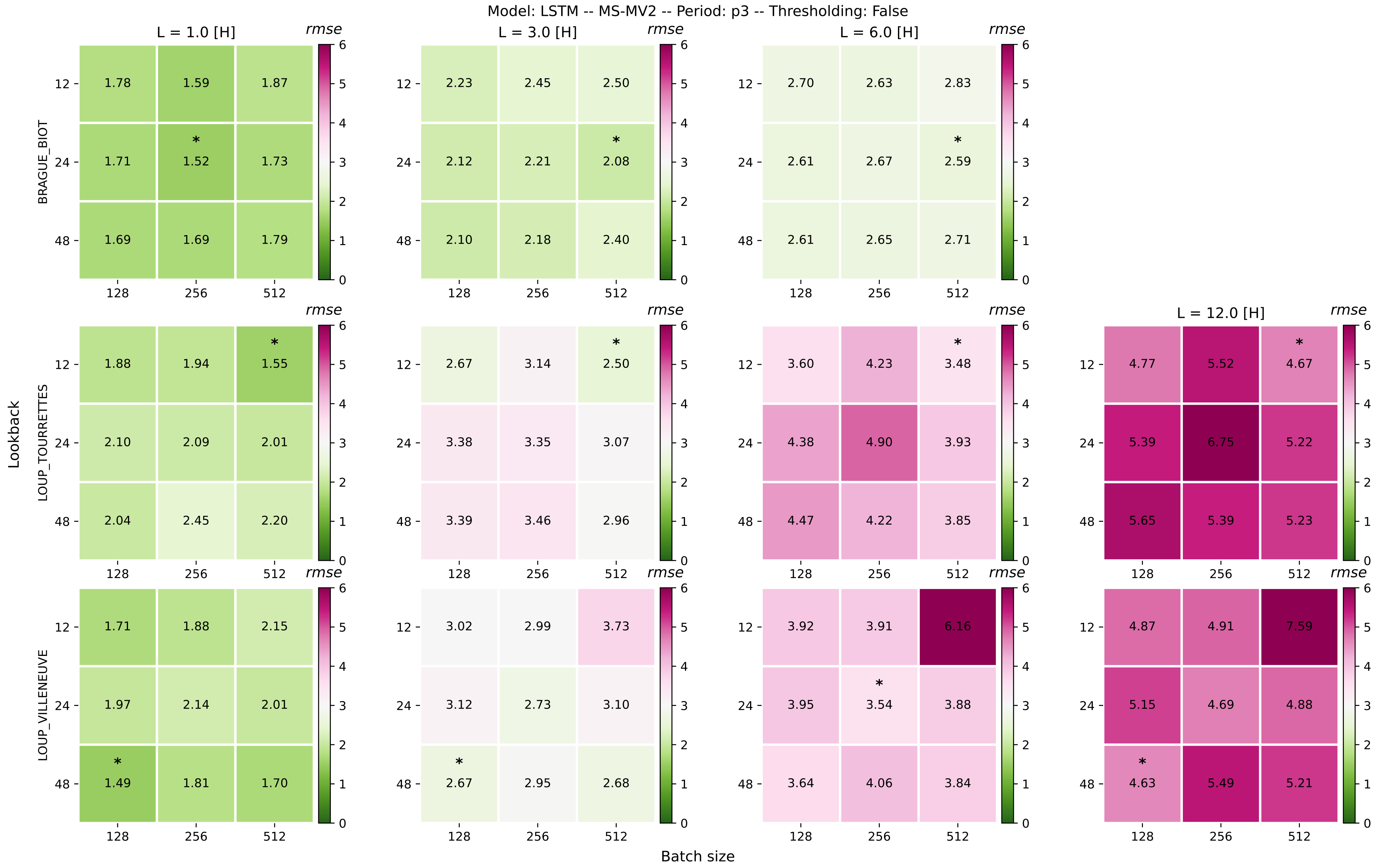

Figure 16: Results of the MS-MV2 model when thresholding is False.

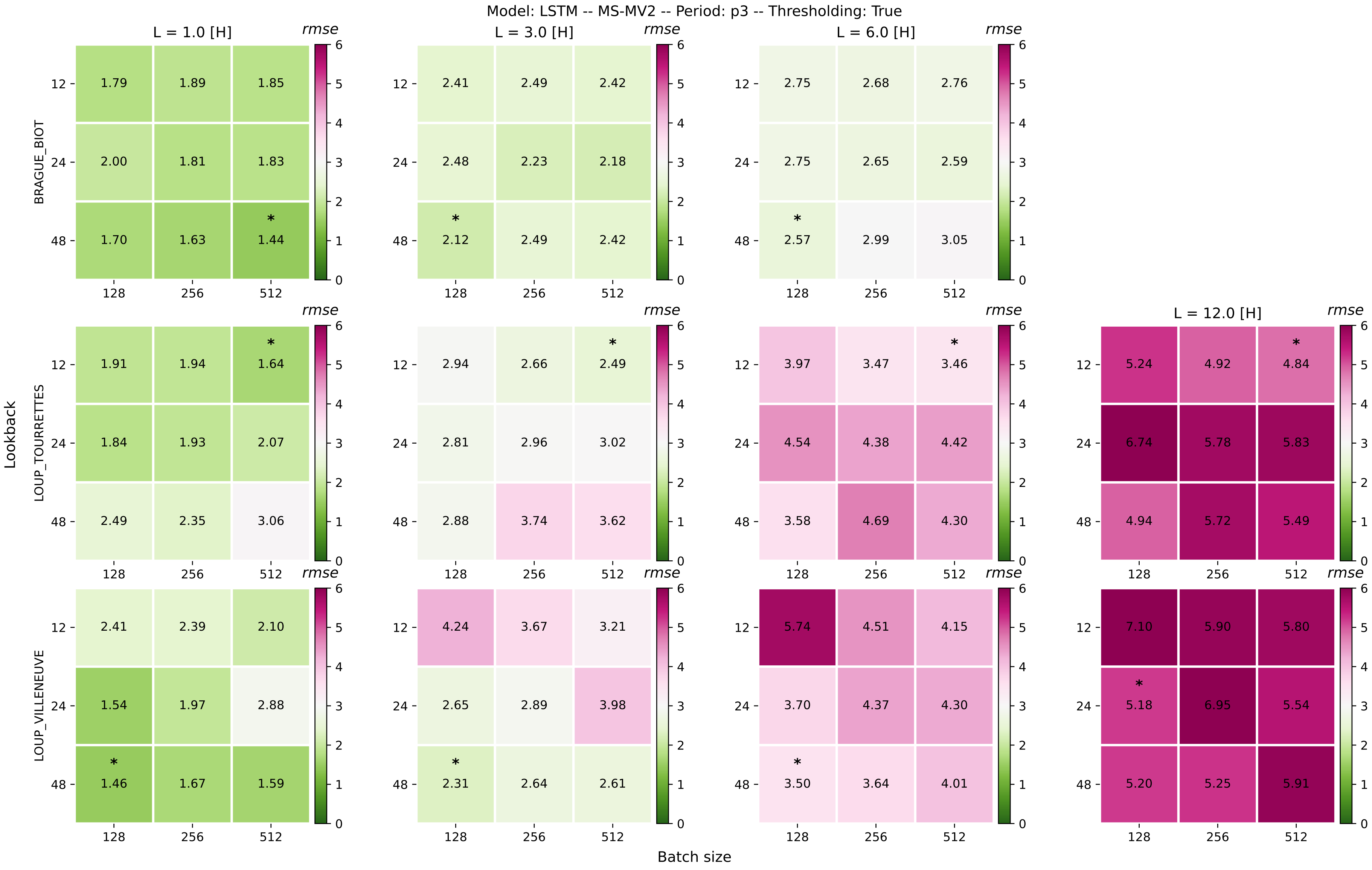

Figure 17: Results of the MS-MV2 model when thresholding is True.

B The Naive model

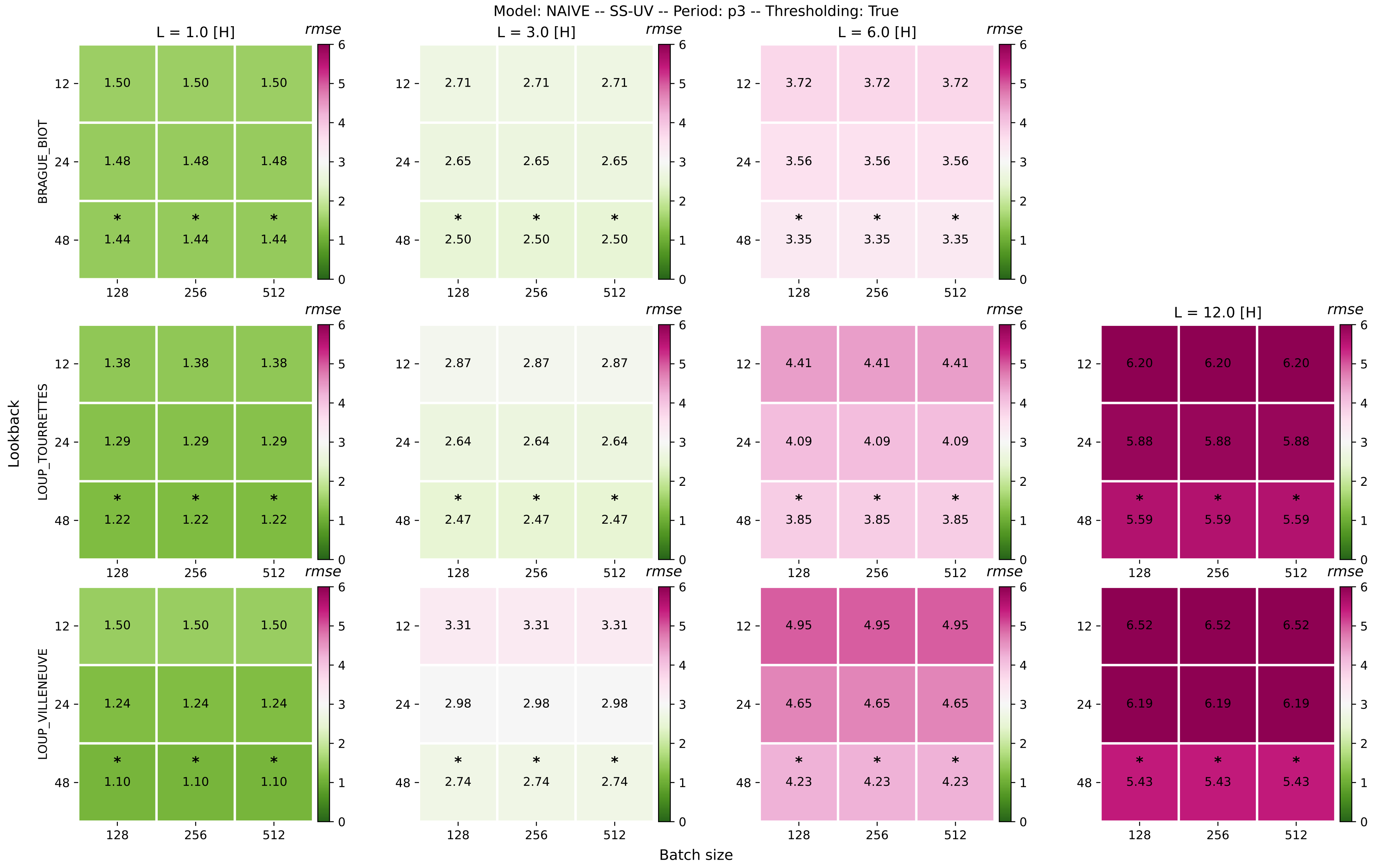

Figure 18: Results of the Naive model for the time steps defined by the SS models.

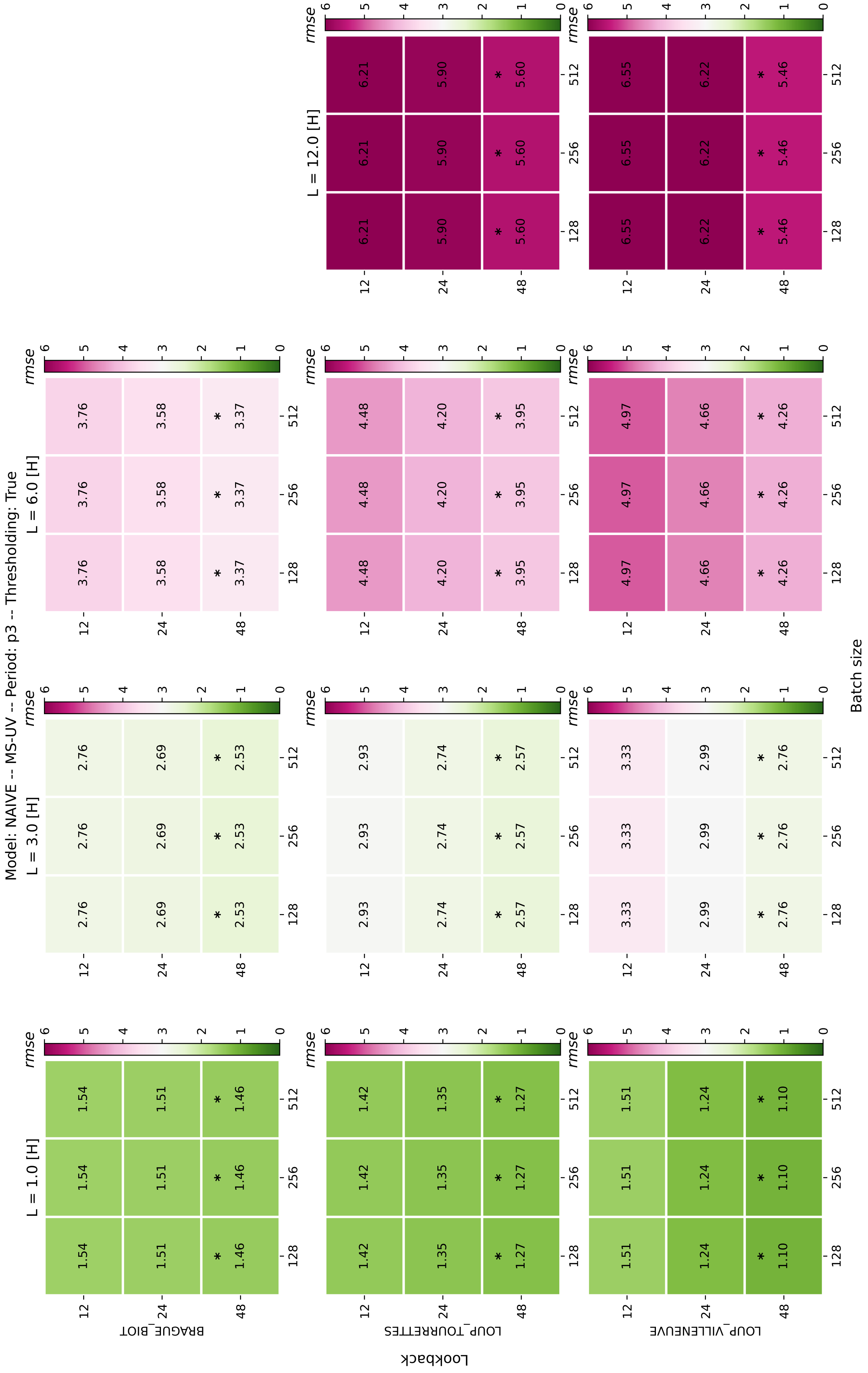

Figure 19: Results of the Naive model for the time steps defined by the MS models.

C The initial GRD model

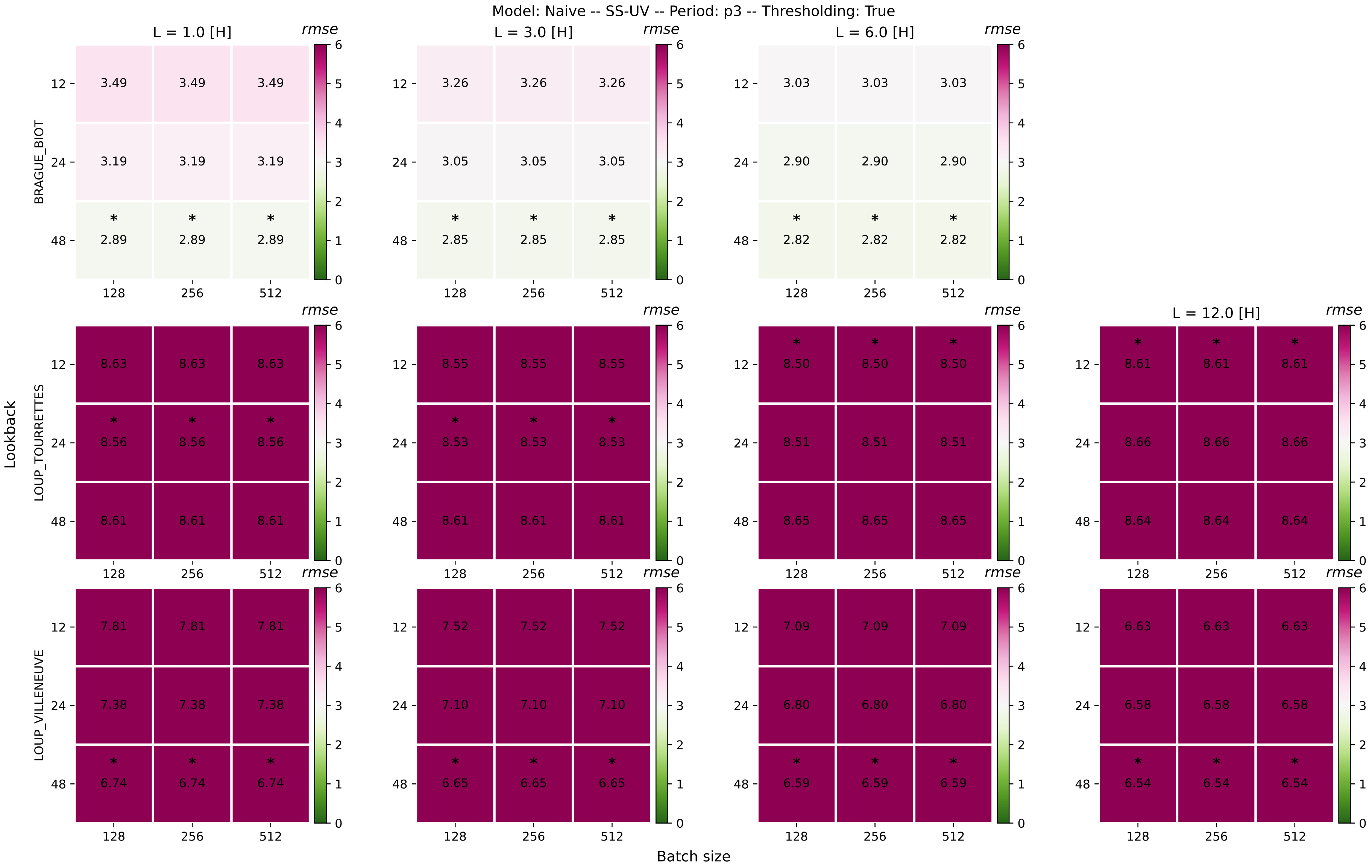

Figure 20: Initial results of the GRD model for the time steps defined by the SS models.

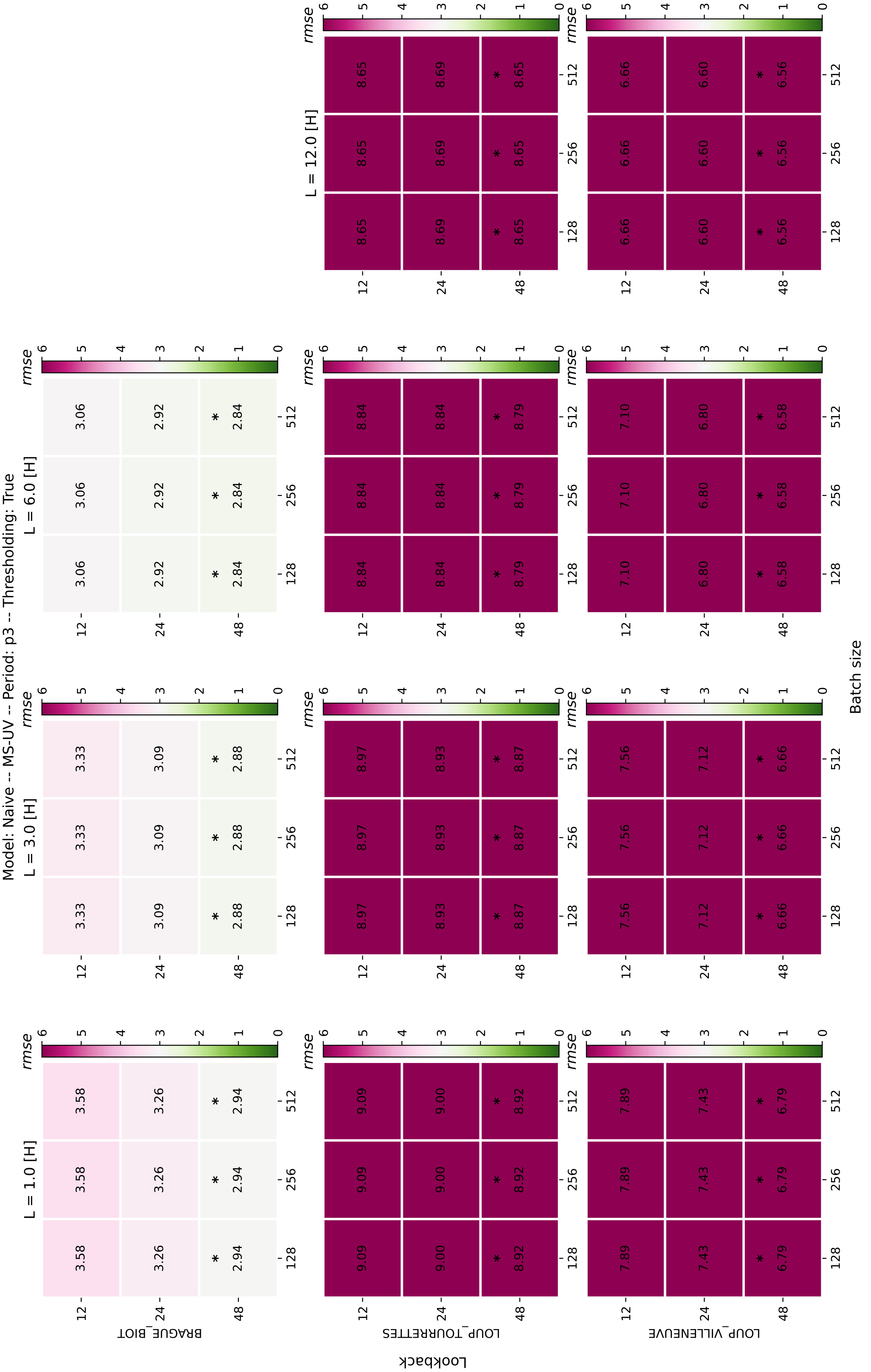

Figure 21: Initial results of the GRD model for the time steps defined by the MS models.

Centre siège Antony-Paris
Direction de l'Appui aux Politiques Publiques
1, rue Pierre-Gilles de Gennes
CS 10030
92761 Antony Cedex
Tél. : +33 1 (0)1 40 96 61 21

Rejoignez-nous sur :
www.inrae.fr

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAE