

深層学習による流量欠測値の補完方法の検討

INVESTIGATION OF MISSING RIVER DISCHARGE DATA IMPUTATION METHOD USING DEEP LEARNING

児島利治¹・Weilisi²・大橋慶介³

Toshiharu KOJIMA, Weilisi and Keisuke OHASHI

¹正会員 博士(工) 岐阜大学准教授 流域圏科学研究センター (〒501-1193 岐阜県岐阜市柳戸1-1)

²Non-Member of JSCE, M. Eng., Graduate school of civil engineering, Gifu University (Yanagido, Gifu, 501-1193).

³正会員 博士(工) 岐阜大学助教 工学部社会基盤工学科 (〒501-1193 岐阜県岐阜市柳戸1-1)

Deep learning has made remarkable progress in recent years, and has many applications in the field of river technology. However, it is hard to say that sufficient knowledge has been obtained for each application target. In this study, Recurrent Neural Network is applied to complement missing values of river discharge and verified. Stochastic gradient descent (SGD) is generally used as the optimization algorithm for deep learning. SGD can obtain good learning results quickly, but the results are slightly different for each learning by its randomized algorithm. The authors performed 1,000 learning and recognized non-ignorable differences in each learning result. The average hydrograph, which obtained by more than 20 learning results, indicates 0.904 in 95 percentile of Nash-Sutcliffe efficiency. The method to reduce the variation of learning results and obtain sufficient accuracy is proposed.

Key Words : RNN, LSTM, missing data imputation, overtraining, stochastic gradient descent

1. はじめに

深層学習, 人工ニューラルネットワーク (Artificial Neural Network; ANN) は近年著しい発展をとげ, 河川技術分野においても多くの適用例がある. 入力層, 中間層, 出力層を持つ3層の階層型ニューラルネットワーク, または浅層型ANN等とも呼ばれるタイプが従来よく用いられてきたが^{1,2)}, より大量の計算資源が利用可能となった近年では, 中間層を多層化した深層型ANN (DNN) が利用されるようになってきた. 和田ら³⁾は, 中間層を5層としたDNNに対し, 遺伝的アルゴリズムによる入力層の構造を決定するGPNNを提案し長期流量予測を行っている. 一言ら^{4,5)}や房前, 島本⁶⁾も同じく中間層を2層とした4層DNNによる洪水予測を行っている. また, 木村ら⁷⁾による排水機場遊水地の水位予測への適用, 田村ら⁸⁾によるダム流入量の長期予測への適用と様々な適用事例がある. 現在のいわゆる第3次AIブームでは, 比較的構造が単純な階層型からより複雑な構造のANNが数多く利用されている. 畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) は, 2次元の画像データを入力データとし, 画像の畳み込みに相当する畳

み込み層, 畳み込み層で抽出された特徴が平行移動などによる影響を受けないように処理するプーリング層を持ち, 顔認識や自動車の自動運転等の画像認識, 動画認識の分野で数多くの適用例がある. 2次元データを用いるCNNは時系列データを利用する河川分野での適用例は少ないが, 1次元データを2次元に変換して利用することも可能である. Kimuraらは, 降雨や水位等の16の変数を入力データとし, 上流から下流に並べた観測点を縦軸に時系列を横軸として2次元データに変換してCNNに適用し, 洪水予測に用いている⁹⁾. 一方, 再帰型ニューラルネットワーク (Recurrent Neural Network; RNN) は時系列データの解析を行うため, 中間層や出力層の状態を一時的に保存し, 次の入力時に利用する構造を持っており, 時系列データの利用に適したANNとして機械翻訳や音声認識によく利用されている. 河川分野においても, 例えば出力層の状態を中間層へ利用するJordan型と, 中間層の状態を同じく中間層へ利用するElman型を組み合わせ, 洪水時の流出予測に適用した寺西, 四俵による研究¹⁰⁾や, ダム流入量予測に適用した谷口らの研究¹¹⁾, Elman型ネットワークを中間層と出力層の両方に適用したRNNを実時間下水道水位予測に適用した研究¹²⁾, 階層型ニューラルネットワークとRNNによる河川流量の予

測精度を比較した研究¹³⁾等がある。しかし、Elman型やJordan型のような単純なRNNでは、勾配消失問題が原因で非常に古い情報を考慮した学習ができないという問題がある。この勾配消失問題を軽減させるため、ゲートと呼ばれる情報の取捨選択機構をもったLSTM (Long Short-Term Memory) やGRU (Gated Recurrent Unit)といったRNNモデルが提案されており、河川分野ではLSTMの適用事例が比較的多い^{14),15),16)}。深層学習に用いられる様々なアルゴリズムはPythonやMATLAB等のプログラミング言語においてパッケージ化、モジュール化され、初心者でも比較的容易に導入する事が可能となっている反面、良好な結果を得るためには、適切なモデルの選択、細かいチューニングといった試行錯誤を必要としており、今後も適用対象毎に多数の解析を実施し、様々な知見を積み上げていく必要がある。

河川水位・河川流量は、河川工事、河川管理における最も基礎的かつ重要な水文情報であり、多くの計測データが存在する。特に河川水位は流量に対して定常的な計測が容易であり、多数の地点で計測されている。川の防災情報¹⁷⁾は、雨量、水位等の情報をテレメータ等により配信しているが、2012年の国・府県管理地点における観測種別毎の欠測・未受信、異常値の発生割合は水位63.0%、雨量8.9%、水質11.3%と水位データの欠測・未受信、異常値の発生割合が多いことが報告されており、水位データの欠測補完は河川管理においても非常に重要な問題と言える。欠測・未受信、異常値の発生要因としては、雨量では通信経路以上が約40%と最も多いのに対し、水位では、計測装置の故障や水位計の測定下限値より水位が低下したことによる欠測が全体の70%以上であると報告されている¹⁸⁾。水位・流量データの欠測値の補完には、欠測期間が短い場合は前後の観測データを用いた方法が用いられ、欠測期間が長期にわたる場合は、タンクモデル等を用いた方法や、同一河川上の隣接観測所の水位データを用いた補完も行われる。油谷ら¹⁹⁾は、欠測区間近傍の観測値を用いた線形補間、6-point scheme及びタンクモデルによる欠測値補間方法の比較を行い、タンクモデルは概ね全ての状況の欠測を補完できるが、多大な労力と時間がかかる、線形補間、6-point schemeは、洪水時の補間は難しく、平水時は労力が少ないためタンクモデルより適しているのではないかと報告している。またパターン分類とファジィ推論を用いた欠測値補完法も提案されている²⁰⁾。欠測期間が短い場合は油谷ら¹⁹⁾のように欠測期間前後の観測値を用いた補間が有用であること、また上下流の別地点の観測値を利用した補完が比較的高精度であることから、深層学習における適用事例が多い洪水予測や長期水位予測に対して欠測補完の適用事例はほとんど無い。深層学習における水位・流量の予測と欠測補完では、i) 洪水予測は数時間先程度を対象とするが欠測補完は欠測期間に依存するため、1日~数日先の推定が対象となる、ii) 欠測値補完では未来の観測データを

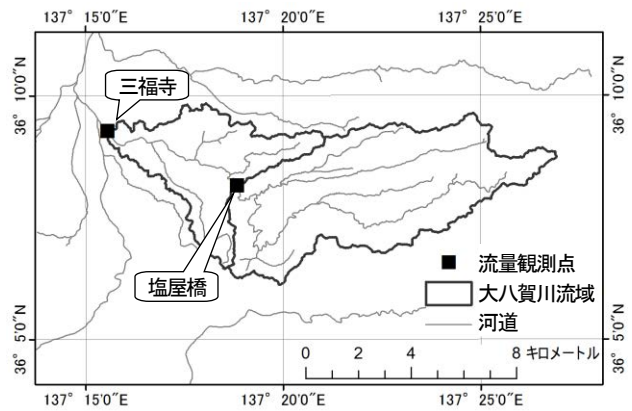


図-1 対象流域。

利用可能である、iii) 欠測値補完では推定対象地点とは別地点であれば、推定対象時刻の観測値を利用可能である事など点で適用方法に違いがある。補完には、周囲の値を用いて間を補う「補間 (interpolation)」と、小尻ら²⁰⁾が用いたように、足りない箇所を補って完全なものにするという意味の「補完 (imputation)」がある。本研究では、河川流量欠測値を、別地点の流量も利用して補完し、完全な流量時系列データを入手しようとするを目的とし、時系列情報の処理に適したLSTMを適用と、その精度検証を行う。

2. 対象流域と使用データ及び線形回帰分析による補完

(1) 対象流域と使用データ

岐阜県高山市の宮川水系大八賀川流域を対象流域とする。大八賀川流域では、岐阜県宮川上流河川開発工事事務所によって、三福寺(集水面積60km²)、塩屋橋(集水面積41km²)の2地点で1986年から1時間間隔で水位・流量の計測が実施されている。本研究では、両地点で欠測の無い2008年の流量観測データで学習、2009年の流量観測データで検証を実施する。図-1に対象流域の図を示す。

(2) 線形回帰分析による補完

図-2に2008年の三福寺流量 $Q_{san}(m^3/s)$ と塩屋橋流量 $Q_{sio}(m^3/s)$ の散布図を示す。2地点間の距離は6.4kmと比較的近距離であり、2地点の流量の相関も相関係数で0.915と非に高い。一方でピーク時には若干のばらつきがあり、降雨イベント毎に違いが見られる。この散布図より線形回帰直線を求めると以下の式が得られる。

$$Q_{sio} = 0.7280 Q_{san} + 0.1657 \quad (1)$$

式(1)を用いて、2009年の Q_{sio} を Q_{san} より推定すると図-3に示すハイドログラフが得られた。ここで横軸はDOY(day

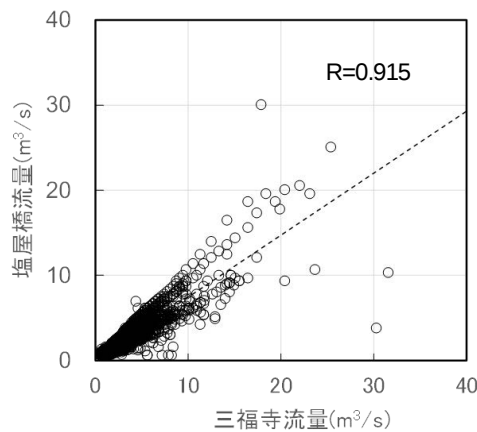


図-2 三福寺と塩屋橋流量の関係。

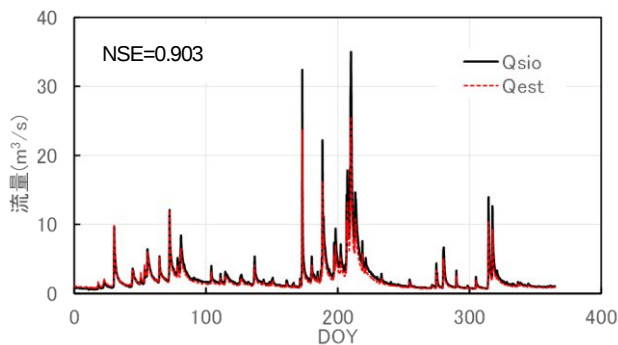


図-3 線形回帰式により推定した流量と観測値の比較。

of year), Q_{est} は Q_{sio} の推定結果である。Nash-Sutcliffe係数(NSE)を求めると、 $NSE=0.903$ と流出予測としては高い精度で推定されている。しかし、ピーク流量の過小評価及び流量低減時と平水時において若干の過小評価が見られる。本研究ではこれらの過小評価が深層学習で修正できるかどうか検討する。

3. LSTMモデル

(1) LSTMモデル

本研究では、図-4に示すような構造のLSTMモデルを用いた。LSTMでは、時刻 t におけるLSTMユニットの内部状態 C_t と、LSTMユニットからの出力 h_t を次の時刻 $t+1$ のLSTMへ伝達する。LSTMユニットでは、時刻 t の入力値 x_t と h_{t-1} 、 C_{t-1} を用いて時刻 t の出力 h_t を求める。出力層ユニットでは、LSTMからの出力 h_t より出力層からの出力 y_t が求められる。以上のような処理の流れは一般的なRNNと同じものである。LSTMでは隠れ層のLSTMユニット内で x_t と h_{t-1} 、 C_{t-1} といった数値を入力層ゲート、忘却ゲートを通すことにより勾配消失問題を緩和させることができるモデルである^{21),22)}。本研究では、図-4に示すように、入力層のユニット数24、出力層のユニット数5、隠れ層(LSTM層)のユニット数を50とした。モデルの実装は、Python3.7.3, Keras2.2.4, Tensorflow1.13.1を用いて行った。

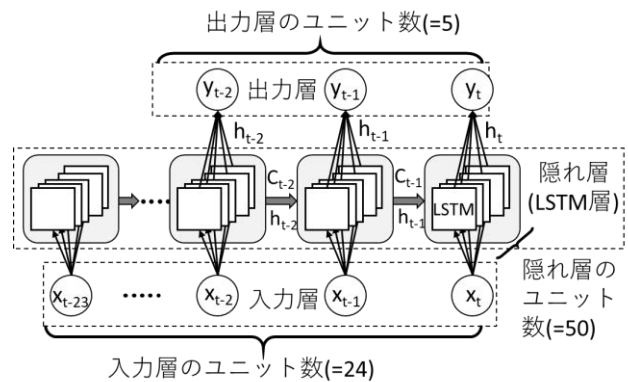


図-4 本研究で用いたLSTMモデルの構造。

表-1 入力データの種類と変数の数。

番号	入力変数	入力変数の数
case1	塩屋橋流量 + 三福寺流量	2
case2	塩屋橋流量	1
case3	三福寺流量	1

LSTM内部の活性化関数、ドロップアウト率等の設定はデフォルトのままとした(すなわち、活性化関数は $\tanh$ 、ドロップアウト率は0)。出力層の活性化関数は elu (Exponential Linear Unit), optimizerはAdamを使用した。損失関数には $\text{mean_squared_error}$ を用いた。学習用データはランダムに分割し、9割を学習データ(Training data), 1割を学習中の検証データ(Validation data)とした。学習回数(epoch数)は最大200とし、過学習防止のため、EarlyStoppingを用いて学習中の検証データを用いた誤差評価(val_loss)の減少が止まってから20回学習を行い停止するよう設定した。

(2) 入力データと出力データ

推定する出力値の時刻を t とすると、出力層では、 $t \sim t-4$ までの塩屋橋流量を出力する。入力層では、 $t \sim t-23$ までの三福寺流量と $t-24 \sim t-47$ までの塩屋橋流量を入力する。時間間隔は全て1時間である。表-1にケース毎の入力データの種類と変数の数を示す。

4. 結果と考察

(1) 学習結果のばらつきに関する検討

ANNの最適化は、一般的に確率的勾配降下法(Stochastic Gradient Descent; SGD)と呼ばれる初期値や学習過程にランダム性を取り入れた手法が用いられるため、学習によって得られるモデル係数は学習毎に少しずつ異なる。図-5にcase1における学習結果を3例示す。この3例は、同じモデル構造、同じ設定、同じ学習用データを用いているが、毎回初期値の設定からリスタートし

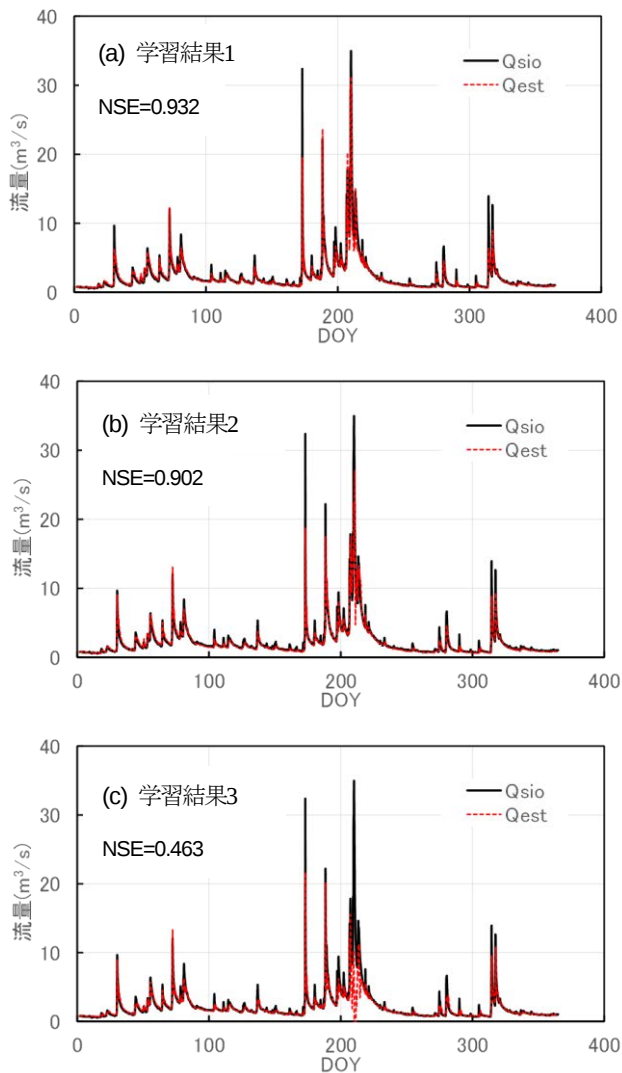


図-5 case1の学習結果の例.

表-2 学習結果ごとの精度評価.

番号	NSE (検証用データ)	loss	val_loss
1	0.932	7.88×10^{-5}	2.40×10^{-5}
2	0.902	8.70×10^{-5}	1.96×10^{-5}
3	0.463	6.80×10^{-5}	2.15×10^{-5}

ており、学習結果が少しずつ異なる。図-5(a)及び(b)の学習結果1, 2は検証用データを用いた精度評価でNSEがそれぞれ0.932, 0.902と線形回帰モデルによる推定結果と同等以上の精度を示し、ピーク流量や流量低減時及び平水時の過小評価が改善されているが、図-5(c)の学習結果3では、NSE=0.463と特にピーク時の再現性が悪い推定結果となっており、学習結果にばらつきが生じる事が分かる。学習結果1～3のTraining dataによる損失関数(loss)とValidation dataによる損失関数(val_loss)の評価値は、表-2に示すようにほとんど同じ値であり、学習時に学習データ以外の適用性、汎用性の確認は、val_lossによる確認はしているものの、困難である。ここで、学

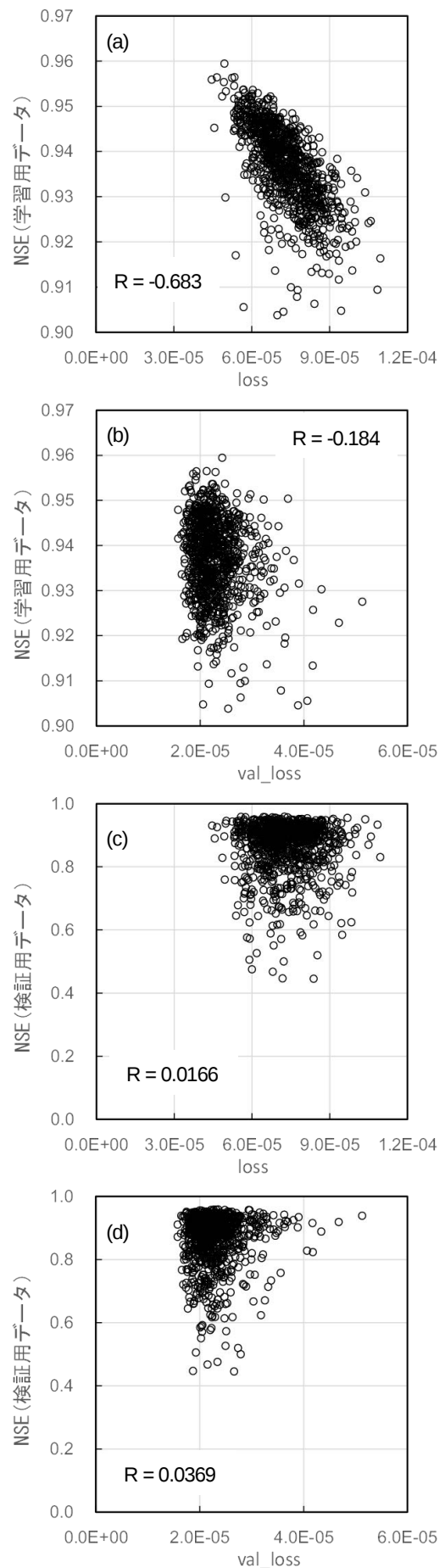


図-6 学習結果の精度評価値の比較.

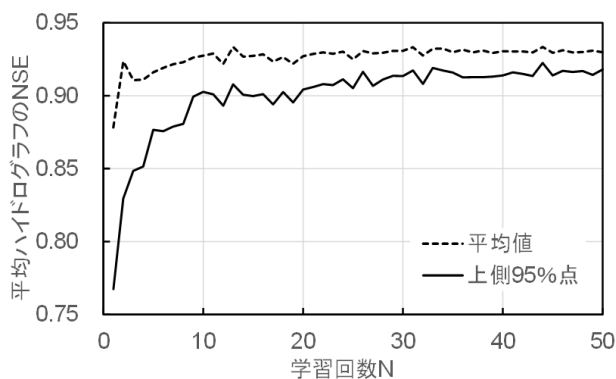


図-7 学習回数とNSEの関係(case1).

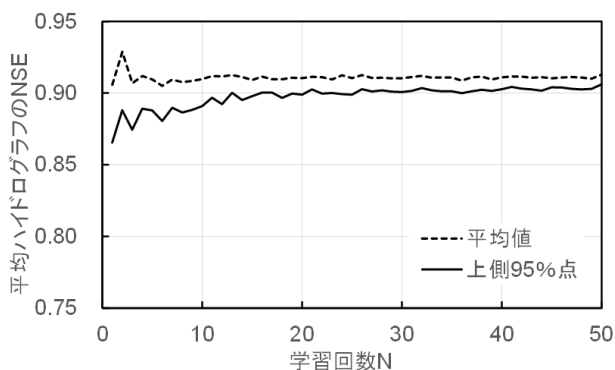


図-8 学習回数とNSEの関係(case2).

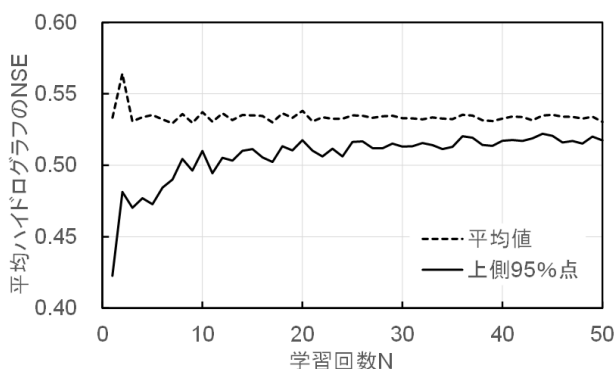


図-9 学習回数とNSEの関係(case3).

学習結果のばらつきがどのように生じているか、独立に1000回学習を行い、学習時のloss, val_lossと学習用データを用いたNSE, 検証用データを用いたNSEを比較した(図-6)。学習用データによるNSEと学習用データの90%を占めるTraining dataによる誤差評価値であるlossの間の相関は $R=-0.683$ と比較的高いものの(図-6(a)), それ以外(図-6(b), (c), (d))の相関は低く、特に検証用データを用いたNSEとloss, val_loss間の相関(図-6(c), (d))はそれぞれ -0.0166 , 0.0369 とほぼ無相関であり、学習時の誤差評価結果が良好でも、適用時の精度が非常に悪い場合が多々ある事が示された。

次に、1000個の学習結果からランダムに N 個抽出した学習結果を用いてハイドログラフを推定し、 N 個のハイ

ドログラフの平均ハイドログラフの精度を評価した。算出する平均ハイドログラフの数を40として、 $N=1\sim50$ までの平均ハイドログラフのNSEの平均値と標準偏差を求め、上側95%点の変化を調べた(図-7)。NSEの平均値では、 $N\geq 2$ の平均ハイドログラフで線形回帰式の精度0.903を超えているが、上側95%点では0.829と線形回帰式の精度を下回る場合がまだまだ多い。上側95%点のNSEは、 $N=1$ の0.768から $N=10$ の0.903と急激に増加し、その後少し増減した後、 $N\geq 20$ でNSE=0.904を常に超過する精度を示すようになった。以上よりcase1では、20回以上の学習結果の平均ハイドログラフを算出することにより学習結果のばらつきを抑え、線形回帰式の精度をほぼ上回る事が分かった。

(2) 入力データの違いによる検討

case2, case3についてもcase1と同様に、学習回数と平均ハイドログラフのNSEの関係を評価した。case2では、case1と異なり、上側95%点のNSEは、 $N=1$ で0.865と学習回数が少ない場合でも高い精度を示したが、 $N\geq 15$ において約0.901でほぼ一定であった。平均値では、常に0.905以上の精度を示しており、50%の確率で線形回帰式を超える精度を示す事が分かった(図-8)。case3では、平均値でも0.530と非常に精度が低かった。上側95%点では N が増加するに従い精度が向上し、 $N\geq 30$ でほぼ一定となるが、その値は約0.52であり十分な精度は得られなかった(図-9)。以上より、欠測値の推定対象の時系列データと隣接観測点の時系列データの両方を用いたcase1が最も推定精度が高く、比較対象である線形回帰式を上回る精度を示す事ができた。case1が最も精度が良かった理由としては、case2は隣接観測点である三福寺の時系列データのみを用いており、直前の塩屋橋流量による補正が行われない事、case3は推定対象の塩屋橋の時系列データのみを用いており、24時間前までの情報では次に流量が増加するのか減少するのかを推定することが難しい事等が考えられる。

5. おわりに

本研究では、時系列データの解析を得意する再帰型ニューラルネットワークに属するLSTMを用いた流量欠測値の補完方法の検討を行った。近年ニューラルネットワークの最適化に用いられるランダム性を含む確率的勾配降下法では、平均的に良い学習結果を高速で得ることができ一方で、得られる学習結果にばらつきが生じる。一般的にその差異は無視しえるほど小さいとされているが、適用対象によっては学習時の精度は高いが別のデータに適用した際に十分な精度が得られない場合がある。この現象はいわゆる過学習と呼ばれるものであり、ANNでは1回の学習に要する計算時間が非常に長いこと

から、学習用データの吟味、モデルの構成、ハイパーパラメタのチューニング等により、汎用性の高いANNを構築する事が一般的な対応法である。一方、本研究の適応対象では、1回の学習に要する時間が1~10分程度と非常に短時間であるため、大量の学習の実施と学習結果のばらつきについて検討した。1000回の学習を行ったところ、無視し得ないばらつきがある事が示唆された。しかし、20回以上の学習結果の平均ハイドログラフは、ほぼ常に比較対象である線形回帰式より良好な精度が得られる事が示され、ハイパーパラメタのチューニング等の詳細な試行錯誤を多数実施せずとも、十分な推定精度を得る方法が提案できた。また、流量欠測値の補完という適用対象においては、精推定対象地点の欠測期間を除く流量データの他に、隣接する別の観測点の流量データを入力データとする事が推定精度の向上に非常に有効である事が示された。

参考文献

- 1) 稲吉昭男, 長江幸平, 民や睦雄, 宮田達磨, 眞間修一, 竹村仁志, ニューラルネットワークモデルによる二級河川での洪水予測の基礎的検討, 河川技術論文集, Vol. 9, pp. 179-184, 2003.
- 2) 阿部清明, 菊池英明, 古川浩平, 塩月善晴, ニューラルネットワークモデルによる流出解析手法 (日流量) に関する研究, 土木学会論文集, No. 656/II-52, pp. 1-13, 2000.
- 3) 和田健太郎, 野原大督, 小尻利治, 関井勝善, 気象情報と人工知能手法を利用した長期流量予測に関する研究, 水文・水資源学会研究発表要旨集, pp.66-67, 2006.
- 4) 一言正之, 櫻庭雅明, 清雄一: 深層学習を用いた河川水位予測手法の開発, 土木学会論文集B1(水工学), Vol. 72, No. 4, I_187-I_192, 2016.
- 5) 一言正之, 桜庭雅明: 深層ニューラルネットワークと分布型モデルを組み合わせたハイブリッド河川水位予測手法, 土木学会論文集B1(水工学), Vol. 73, No.1, pp. 22-33, 2017.
- 6) 房前和明, 島本卓三, AIを活用した洪水予測技術の開発について, 平成30年度九州国土交通研究会論文 (2020/3/5確認 [http://www.qsr.mlit.go.jp/useful/n-shiryo/kikaku/kenkyu/h30/04/4_03\(18\).pdf](http://www.qsr.mlit.go.jp/useful/n-shiryo/kikaku/kenkyu/h30/04/4_03(18).pdf)), 2019.
- 7) 木村延明, 中田達, 安瀬池一作, 関島建志, 桐博英, 馬場大地, 人工ニューラルネットワークモデルを利用した排水機場遊水地の水位予測に関する研究, 農研機構報告, 農村工学部門, Vol. 3, pp. 71-80, 2019.
- 8) 田村和則, 加納茂紀, 三浦心, 山脇正嗣, 金子拓史, ダム流入量長時間予測への深層学習の適用ーダム防災操作の効率化を目指してー, 土木学会論文集B1(水工学), Vol. 74, No. 5, I_1327 - I_1332, 2018.
- 9) Kimura, N., Yoshinaga, I., Sekijima, K., Azechi, I. and Baba, D., Convolutional Neural Network Coupled with a Transfer-Learning Approach for Time-Series Flood Predictions, *Water*, Vol. 12, No. 1, 96, 2020.
- 10) 寺西隆司, 四俣正俊, 流出解析用ニューラルネットワークにおけるリカレント構造の考察, 愛知工業大学研究報告, Vol. 32B, pp.89-96, 1997.
- 11) 谷口純一, 小島崇, 曾田康秀, 福元秀一郎, 佐藤尚, 町田佳隆, 見上哲章, 永山正典, 錦織俊之, 渡邊明英, Recurrent Neural Networkによるダム流入量の予測, 河川技術論文集, Vol. 25, 2019.
- 12) Chiang, Y., Chang, L., Tsai, M., Wang, Y. and Chang, F., Dynamic Neural Networks for Real-Time Water Level Predictions of Sewerage System-covering gauged and ungagged sites, *Hydrology and Earth System Sciences*, Vol. 14, pp. 1309-1319, 2010.
- 13) Sahoo, A., Samantaray, S. and Ghose, D. K., Stream Flow Forecasting in Mahanadi River Basin using Artificial Neural Networks, *Procedia Computer Science*, Vol. 157, pp. 168-174, 2019.
- 14) Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Hermegger, M., Rainfall-Runoff Modelling using Long Short-Term Memory (LSTM) Networks, *Hydrology and Earth System Sciences*, Vol. 22, pp. 6005-6022, 2018.
- 15) Hu, C., Wu, Q., Li, H., Jian, S., Li, N. and Lou, Z., Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation, *Water*, Vol. 10, 1543, 2018.
- 16) Le, X., Ho, H. V., Lee, G. and Jung, S., Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting, *Water*, Vol. 11, 1387, 2019.
- 17) 国土交通省HP「川の防災情報」 <http://www.river.go.jp/kawabou/> (2020/1/24確認)
- 18) 小高茂治, 近畿地方整備局における河川情報の監視と水文観測データの品質照査について, 平成25年度近畿地方整備局研究発表会論文集, 2013.
- 19) 油谷涼, 大串浩一郎, 手塚公裕, 有明海流入河川の流量欠測データの補間方法について, 土木学会西部支部研究発表会要旨集, pp. 155-156, 2011.
- 20) 小尻利治, U.S.PANU, 友杉邦雄, パターン分類とファジイ推論による欠測流量データの補完, 水文・水資源学会誌, Vol. 7, No. 6, pp. 536-543, 1994.
- 21) Hochreiter, S. and Schmidhuber, J., Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.
- 22) Gers, F. A., Schmidhuber, J. and Cummins, F., Learning to Forget: Continual Prediction with LSTM, *Neural Computation*, Vol. 12, No. 10, pp. 2451-2471, 2000.

(2020. 4. 2受付)