RIVER-D-23-00095

転移学習を用いた洪水イベントAI事前学習モデルの構築

○木村延明[1]・皆川裕樹[1]・福重雄大[1]・馬場大地[2]（1 農研機構農工部門，2 ㈱アーク情報システム）

課題設定

大量のデータさえあれば予測可能なニューラルネットワーク（ANN）は、迅速性・簡便性の利点から、近年河川洪水予測に利用されるものの、過去のデータに含まれない（学習されていない）洪水事象の予測は困難である。特に、極端気象の影響が懸念される昨今、未経験の洪水事象が発生する可能性があり、ANNの利用範囲を拡張することが求められる。

課題解決のための仮説

予測対象流域の洪水事象のデータが少なければ、以下のようにデータを割り増しする。
１）対象流域に最適化された物理モデルなどを用いて、仮想データを生成する。
２）他流域のデータを掻き集めて転用する。
　本研究では、２）を採用し、他流域のデータの特徴のみ抽出した事前学習モデルを生成し、それを予測対象流域で再利用し、局所最適化するために、転移学習を用いる。

研究方法と成果

水文環境などが類似する九州地方の複数流域で観測された水位データ（但し、予測対象地点のデータ除く）をまとめて、ANNを用いてAI事前学習モデルを生成し、予測対象データで再学習を行い、洪水時の水位予測を行った（図1）。従来型ANNとの比較結果では、対象ドメインデータが多い地点（図2左）、少ない地点（図2右）で誤差評価に基づけば、前者は最大3%、後者は最大70％の改善が見られ、AI事前学習モデルを少ないデータに適用した場合、高精度な予測結果が得られた。

具体的データ

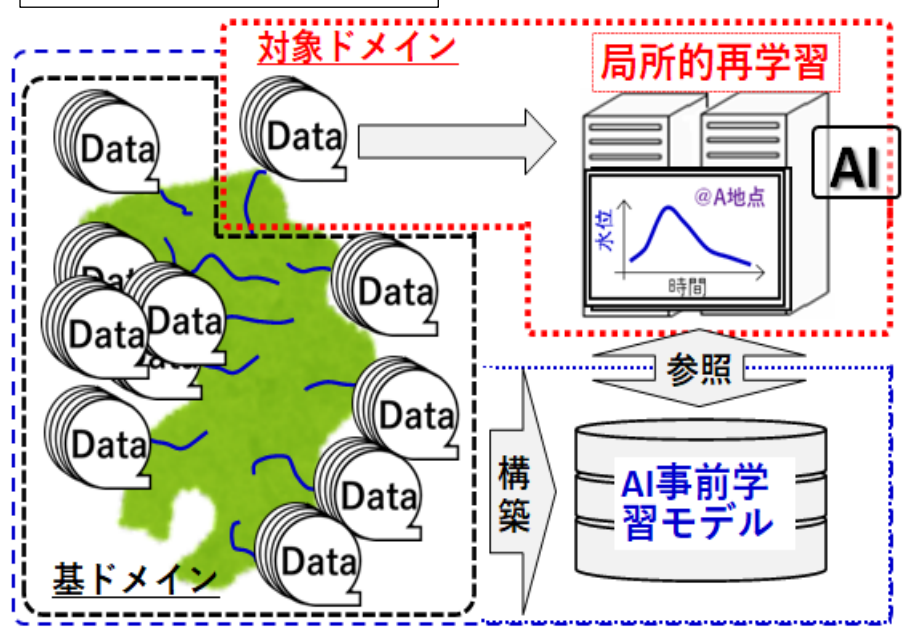


図1　提案する予測手法のイメージ

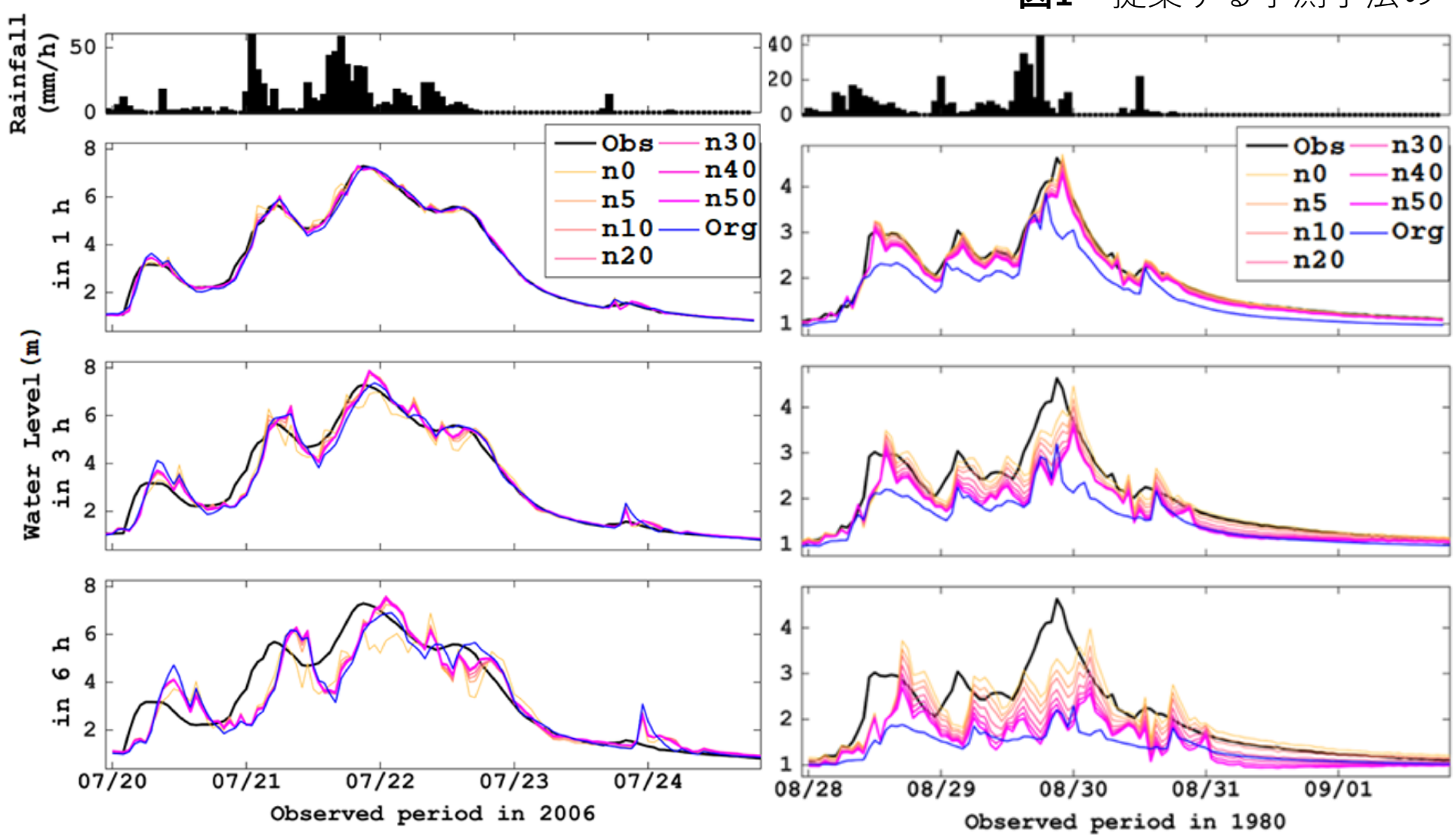


図2 降雨量と新・従来型ANNのリードタイム1 h, 3 h, 6 hの水位予測結果と観測値の比較，黒線＝観測値，青線＝従来型ANN，黄色～ピンク＝新ANN，n＝再学習回数.

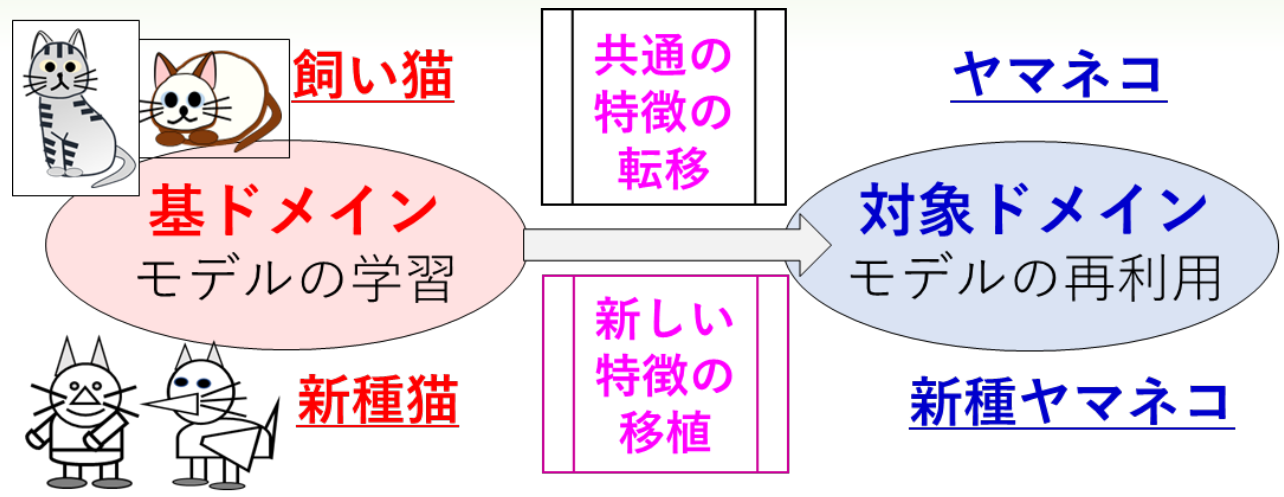
課題と今後の方向性

両地点ともにリードタイム6 hでは、LSTMの構造に起因するタイムラグの発生で、観測値の再現性が低下した。今後の改善が求められる。

RIVER-D-23-00095

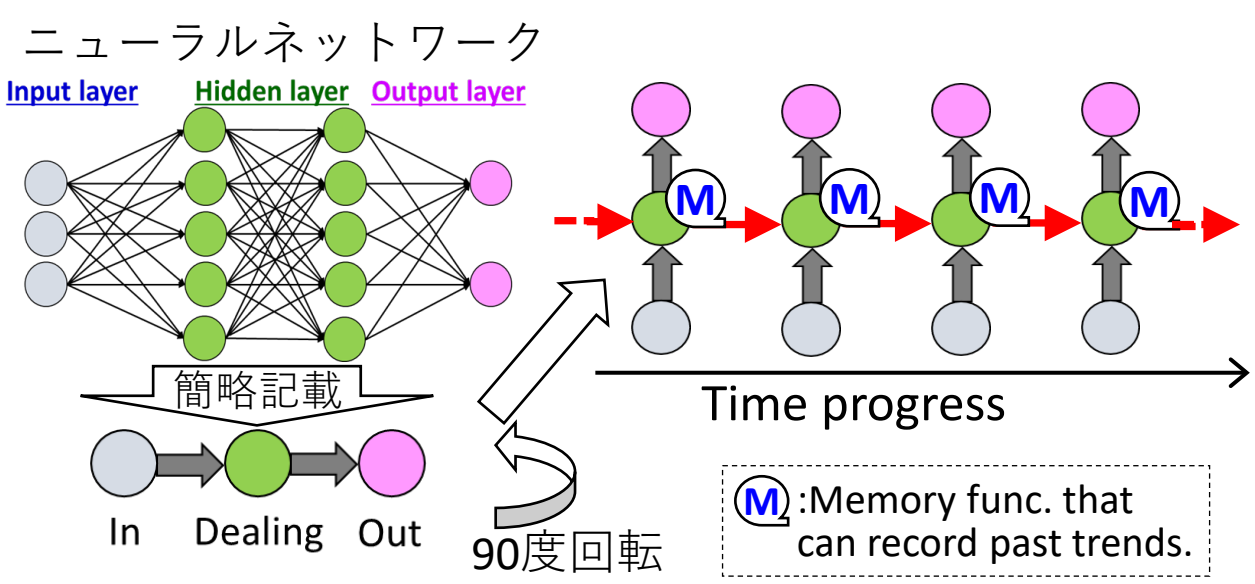
補足説明用資料

（機械学習における）転移学習とは、大量の学習データを有し、ソースとなるドメイン（基ドメイン）のデータで事前学習されたモデルを、予測対象のドメイン（対象ドメイン）のデータで再利用し、基ドメインデータの特徴を、対象ドメインデータに転移する方法である。例えば、飼い猫の画像の特徴を共通の特徴をもつヤマネコに転移する場合や、新種猫の新しい特徴をヤマネコに移植することで新種ヤマネコを誕生させる場合に利用可能である（右図）なお、本研究ではANNをモデルとして使用した。



転移学習のイメージ図

本研究のANNモデルは、連続データの予測に有用な長・短期記憶（Long short-term memory: LSTM）を採用した。時間ステップの横の情報を伝達する構造になっており、さらに長期・短期のデータの変化を記憶する機能を備えたアーキテクチャである（下図参照）。



LSTM構造のイメージ図

本研究で利用したデーター覧は、11水系の24カ所で収集された、1時間間隔の120ステップの洪水イベントデータである（下表）。対照的な地点である赤丸の予測結果のみを表示した。

予測対象地点などの情報一覧表

水系名：地点名	氾濫危険水位(m)	洪水データ数	流域面積 (km^2)	降雨量観測地点	予測対象地点
大淀川：大田原	9.20	2	101.0	嵐田	
岳下	7.50	9	160.0	岳下	CS0
樋渡	3.20	6	861.0	樋渡	
筑後川：花月	3.35	3	130.2	花月	
杖立	6.00	16	286.0	杖立	CS1
端間	4.66	2	167.0	原田	
菊池川：岩崎	1.70	3	18.2	高瀬	
佐野	6.00	9	156.0	合志	CS2
城	4.60	3	109.8	岳間	
広瀬	3.90	3	152.0	赤星	
袋田	3.90	2	87.9	城北	
川内川：栗野橋	7.00	17	339.0	栗野	○CS3
花北	4.80	6	245.0	大口	
松浦川:徳須恵橋	5.20	10	71.0	畑川内	CS4
肝属川：姶良橋	3.10	5	62.0	大平	CS5
王子橋	5.00	1	46.0	大浦	
俣瀬	5.50	3	450.0	高山	
遠賀川：伊田	5.20	3	127.0	小柳	○CS6
野面	4.00	2	8.0	直方	
大分川：宮苑	2.94	2	53.0	大分	
五ヶ瀬川：松山	5.90	2	1,072.0	延岡	
三ツ瀬	5.30	2	1,053.0	同上	
本明川：裏山	3.70	1	35.8	清水	
山国川：馬場	5.80	2	34.0	馬場	

LSTMモデルのタイムラグが生じる理由（そもそも論）

LSTMのデータ処理構造の問題で発生する。連続データの予測では、他機械学習よりも卓越し、頻繁に利用されるものの、リードタイムが長くなれば、時間遅れの状況が現れる。その理由は、そもそも言語処理の予測モデルから発達したもの、以下の例を示しながら課題を指摘したい。

自己紹介の事例文章「私は、太郎と申します」を考えます。例えば、3つの文節に分割「私は、」「太郎と」「申します。」「私は、」という入力情報が入ってきたとき、「申します。」を予測するのは難しい。何故なら、「私は、」の後は、必ずしも人名が来るとは限らない。例えば、「家に（います。）」と物が続く場合も考えられる。従って、人名の「太郎と」いう文節が続かなければ、「申します。」という文節を選択できない。このことから、より離れた文節を正確に予測するのは困難で、タイムラグのような問題が生じる。