河川技術論文集，第26巻, 2020年6月

ランダム・フォレストを用いた融雪期のダム流入量予測における入力データの検討

EXAMINATION OF INPUT DATA FOR PREDICTION OF DAM INFLOW DURING SNOWMELT SEASON USING RANDOM FOREST

山田嵩[1]・阿部真己[2]・滝口大樹[3]・谷瀬敦[4]・矢部浩規[1]
Takashi YAMADA, Masami ABE, Hiroki TAKIGUTI, Atsushi TANISE and Hiroki YABE

[1]正会員　国立研究開発法人土木研究所　寒地土木研究所
（〒0620-8602 北海道札幌市豊平区平岸1条3-1-34）
[2]正会員　いであ株式会社（〒224-0025 神奈川県横浜市都筑区早渕2-2-2）
[3]正会員　いであ株式会社（〒154-8585 東京都世田谷区駒沢3-15-1）
[4]正会員 国土交通省　北海道開発局札幌開発建設部（〒060-8506 北海道札幌市中央区北2条西19丁目）

In predicting the inflow of dams during the snowmelt season by deep learning, it's important to select input data. In this study, we visualized the importance of input data using the random forest method. The input data includes precipitation, global solar radiation, reflected solar radiation, upward radiation, downward radiation, surface temperature, temperature, wind speed, humidity, snow weight, snow depth and snowmelt amount.

As a result, the most important factors were the temporal distribution of precipitation, upward radiation, air temperature, surface temperature, and snow depth.

Key Words : *random forest, snowmelt, Dam inflow forecast*

1．はじめに

積雪寒冷地においては融雪水をダムに貯留して，春先から初夏にかけての水需要を賄っている．一方で，融雪期においては気温の急上昇や強雨等により，大規模の洪水を引き起こすこともある．そのため，融雪期における高精度なダム流入量予測は，水資源の有効活用や融雪洪水の防止といった観点から極めて重要である．近年では人工知能を活用した研究が水文分野でも進められており，河川の水位予測や夏季のダム流入量予測にて成果を挙げている．一方で，融雪期のダム流入量予測を対象とした研究は少ない．

現在，予測に用いられている人工知能(AI)は主に深層学習を用いたものである．これらの予測モデルは，降雪・積雪・融雪・ダム流入量の複雑なプロセスをブラックボックス化させるため，入力データの質や量が予測結果に多大な影響を与える．そのため，過学習を防止し高精度なモデルを構築するためには，入力データの適切な取捨選択が求められる．しかしながら，深層学習では入力データの重要度を示すことは不可能であり，予測結果から逆算して推定するしかない．滝口ら[1)]の研究では，奈良俣ダムの融雪期の日流入量予測と予測に用いる観測項目の抽出を行い，その結果，日降水量，最新積雪深，日平均風速，日平均気温の過去6日分のデータが重要であるという結果が得られている．しかし，滝口らの研究ではニューラルネットワークを用いた予測結果から重要度を推定しており，入力データの重要度を定量的には表現できない．

本研究では，融雪期の時間単位のダム流入量予測を実施する際に、過学習防止と予測精度向上に必要となる水文・気象データの観測項目の絞り込みを行うため変数重要度を示すことが可能なランダム・フォレスト（以後，RFという）により，入力データの重要度を示した．

2．研究対象

(1) 対象領域

本研究にて対象領域としたのは北海道札幌市に存在す

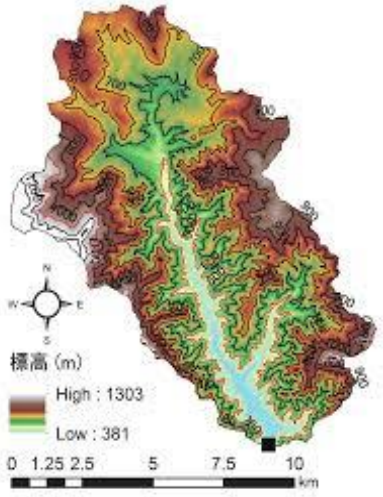

図-1　定山渓ダム流域

表-1　重要度計算の対象としたデータ

観測機関	観測地点	項目	単位
寒地土研	流木処理場	全天日射量	W/m²
寒地土研	流木処理場	上向き放射量	W/m²
寒地土研	流木処理場	下向き放射量	W/m²
寒地土研	流木処理場	反射日射量	W/m²
寒地土研	流木処理場	表面温度	℃
寒地土研	流木処理場	気温	℃
寒地土研	流木処理場	風速	m/s
寒地土研	流木処理場	湿度	%
寒地土研	流木処理場	積雪重量1	kg/m²
寒地土研	流木処理場	積雪重量2	kg/m²
寒地土研	流木処理場	積雪深	cm
寒地土研	流木処理場	融雪量	mm
国土交通省	春香山	積雪深	cm
気象庁		解析雨量 (時空間分布)	mm

る定山渓ダムである．定山渓ダムの流域面積は104 km²であり，定山渓ダムでは寒地土木研究所（以後，寒地土研という）が，全天日射量，反射日射量，上向き放射量，下向き放射量，表面温度，気温，風速，湿度，積雪重量，積雪深，融雪量の観測を行っており，これらのデータを活用してダムへの時間流入量予測手法の構築を検討している．ただし，積雪重量は2ヶ所で観測しており，重要度計算の際には平均値を用いた．流域図を**図-1**に示す．

(2) 解析対象データ

本研究にて重要度計算の対象としたデータを**表-1**に示す．対象期間は2006年から2016年の3月から5月までである．ただし，異常値の除去，短い欠測値の線形補間を行い，数日以上の長期間の欠測した項目がある場合には解析期間から除外している．最終的に用いた解析対象のデータ期間を**表-2**に示す．

表-2　解析に用いるデータの期間

年	日時		データ数	
	開始	終了	時間	日
2006	2006/3/1 0:00	2006/3/18 23:00	432	18
2008	2008/3/1 0:00	2008/5/26 23:00	2088	87
2009	2009/3/1 0:00	2009/5/24 23:00	2040	85
2010	2010/3/1 0:00	2010/5/24 23:00	2040	85
2011	2011/3/1 0:00	2011/5/23 23:00	2016	84
2012	2012/3/1 0:00	2012/5/31 23:00	2208	92
2013	2013/3/1 0:00	2013/5/31 23:00	2208	92
2014	2014/3/1 0:00	2014/5/6 23:00	1608	67
2015	2015/3/18 0:00	2015/5/3 23:00	1128	47
2016	2016/3/1 0:00	2016/5/31 23:00	2208	92

3．解析手法

(1) 前処理

a) 次元圧縮

空間分布，日変動，時間累積の特徴まで，全ての情報を考慮すると，表の多変量データの次元数は，1セットにつき15128次元のデータとなり，非常に高次元のデータとなる．このような高次元のデータでは「次元圧縮」という手法により変数を要約して変数量を削減する前処理が有効である．ここでは，最も基本的な方法として「主成分分析」という手法を用いて，情報量の約95 %を保つのに必要な次元数にまで圧縮した．

主成分分析により圧縮された前後の次元数の比較を**表-3**に示した．当初15128次元であったデータは134次元にまで圧縮された．

b) 降水量の空間分布の圧縮

空間分布の特徴をとらえた次元圧縮は，主成分分析では限界があることから，深層学習の手法の一つであるVAE(Variational Auto-Encoder)[2]という手法を用いて降水量の空間分布は2次元にまで圧縮している．VAEを用いることで，主成分分析などの次元圧縮よりも，より小さい次元で，より高精細な情報を表現することができる．

図-2にVAEの概要を示す．もともとの情報（14×17次元の高次元情報）を一度2次元にまで圧縮された情報に変換して，これをもう一度高次元の情報に復元することを学習する深層学習モデルである．入力と同じものを出力するだけであるので，教師データのアノテーションは不要である．様々な降水の空間分布を小さい次元の数字のベクトルのみ（以後，埋め込み変数という）で表現するが，この埋め込み変数はある決まった分布（ここではガウス分布）となるように近似される．埋め込まれた降水の空間分布をモーフィングと呼ばれる手法で確認した結果、埋め込み変数①が降水量の空間的な偏りをつかさどり，埋め込み変数②は降水量の大小をつかさどっていることが確認できた．

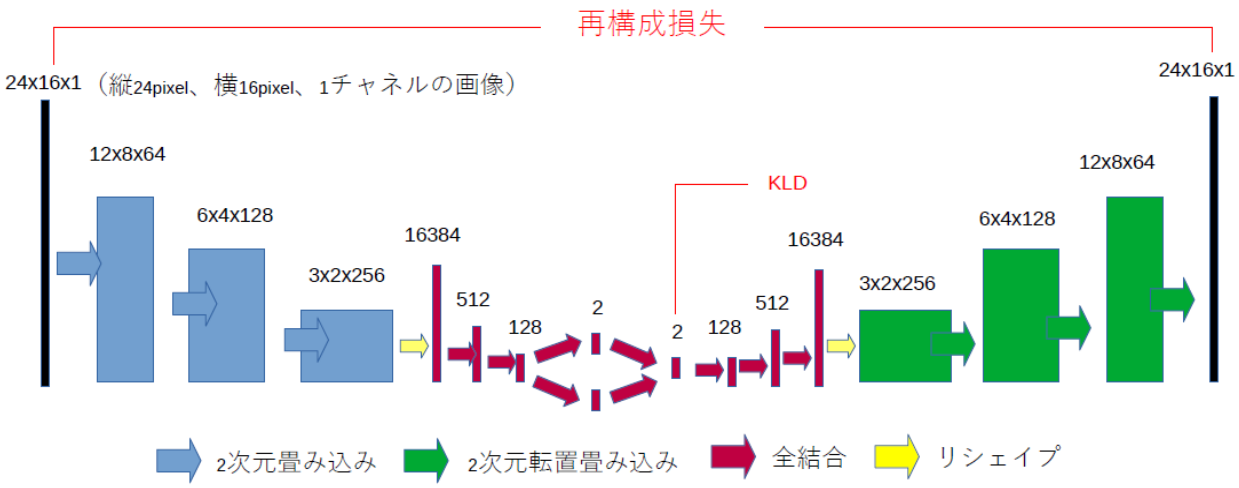

図-2　VAEの概要

表-3　主成分分析によるデータの圧縮

観測機関	観測地点	項目	単位	95%を説明するモード数		本来の次元数	
				当日の時間パターン	日平均の1週間変動パターン	当日の時間パターン	日平均の1週間変動パターン
寒地土研	流木処理場	全天日射量	W/m2	4	6	24	7
寒地土研	流木処理場	放射量(下向き)	W/m2	5	5	24	7
寒地土研	流木処理場	反射日射量	W/m2	3	6	24	7
寒地土研	流木処理場	放射量(上向き)	W/m2	2	2	24	7
寒地土研	流木処理場	表面温度	℃	2	2	24	7
寒地土研	流木処理場	風速	m/s	12	6	24	7
寒地土研	流木処理場	気温	℃	1	3	24	7
寒地土研	流木処理場	湿度	%	6	6	24	7
寒地土研	流木処理場	積雪重量(1,2平均)	kg/m2	1	1	24	7
寒地土研	流木処理場	積雪深	cm	1	1	24	7
寒地土研	流木処理場	融雪量	mm	4	5	24	7
国交省	春香山	積雪深	cm	1	1	24	7
—	—	降水量空間分布（埋め込み①）	mm	18	6	5712	1666
—	—	降水量空間分布（埋め込み②）	mm	18	6	5712	1666
合計			次元数	78	56	11712	3416
				134		15128	

(2) 主成分分析

主成分分析（または経験的直交関数）を用いて1日の時間変動と，日平均データの1週間変動の代表的な変動パターン（因子負荷量，モードのパターン）と各変動パターンの変遷の様子（主成分得点，モード）を抽出する．例えば気温の時間データの場合、0時～23時の24次元の多変量の毎日のデータと考え、1日の代表的なパターンが毎日どのように出現しているかに直交分解することを行う．気温は夜間に低く日中高い値になる日周期の変動が卓越しているため第1モードは日周期変動，それ以降はより細かい周期の変動が抽出される．概ねFFTなどのような周波数分解とよく似た結果となる．

降水の空間分布の埋め込み変数も含めて，様々な変数のそれぞれの日変動パターンを主成分分析でモードを抽出した．週間変動については，全項目を日平均した後に，代表的な週間変動を抽出した．

(3) 流入量の1日のパターンの類型化

RFによる変数の重要度を可視化するにあたり，目的変数の分類データを作成する必要がある．分類の対象としてはダム流入量(m^3/s)の「1日の波形のパターン分類」とした．類型化には，主成分分析により 1日の代表的な変動パターンの変遷（主成分得点）を抽出して，主成分得点の類似度をもとに類型化した．

クラスター分析は，ユークリッド距離で距離を評価しながら，Ward法で階層的クラスリングを行った．これにより代表的な日パターンとして近い・遠いが評価可能で，日パターンそのものの分類が実現できるものと考えられる．本分析の閾値は100とし，その結果1日の波形のパターンは全14パターンに分類された．

(4) RFによる分類と重要度の可視化

RFは決定木（Decision Tree）をアンサンブル的に行う解析手法である．決定木は，多次元の変数のうち，データを最もよく分離する変数とその閾値を順番に選択して，データを当該変数で2分割することを繰り返すことで最終的な分類を得る分類器である．最初に分類に選ばれた変数は，データをうまく分類するために最も重要と考えられる変数であるため，変数の重要度のランキングが可能である．なお，分類の指標としては不平等さを評価するジニ係数やエントロピーなどを用いることが多い．RFはこの決定木をランダムに抽出された複数のデータセットに対してアンサンブル的に行う分析方法である．それぞれの決定木は個別に変数の重要度のランキングの情報を有しているため，変数の重要度のランキングを各決定木の多数決で決めることができる．これが，RFが変数の重要度を可視化できる仕組みである．ランダムに抽出されたデータセットに対して決定木をアンサンブル的に実施（木の集合体はForest（森））することから，「ランダム・フォレスト」と名付けられている．

各モードを対象に，RFによるフィッティングと重要度の可視化を行った．日変動のモードは当日分，週間変動のモードは前日までの1週間分を用いて，図に示した分類の予測を学習させた．フィッティングは全データの20 %を検証用に用いており，学習データの正解率100 %，検証用データの正解率60.5 %であった．

4．結果と考察

(1) ダム流入量パターンの分類化結果

1日の波形のパターン分類の結果を，分類ごとに日変動パターンのデータをすべて可視化したものを，各グループを色で示した各年の時系列グラフを**図-3**から図-6に，ボックスダイヤグラムを**図-7**に示す．ただし，紙面の都合により時系列グラフは一部の偶数年を，ボックスダイヤグラムは奇数番号のパターンのみを示している．似たような日パターンにうまく分類できていることが確認できる．

(2) 各変数のモード抽出結果

主成分分析により各変数のモードを抽出した結果を**図-8**に示す．ここでモードは情報量の累積が95 %の上位を可視化しており，紙面の都合により一部の変数は割愛している．

(3) RFによる重要度の可視化結果と考察

RFによる各変数の全モードの重要度を**表-4**に，各変数の重要度の最大値を**表-5**に，重要度が0.01以上の抽出結果を**表-6**に示す．表-6の●及び○が0.01以上の項目を表し，重要度がより高い項目を●で表記している．気温，

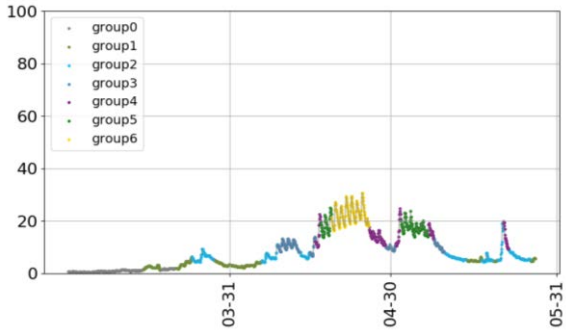

図-3 各グループの時系列グラフ（2008年）

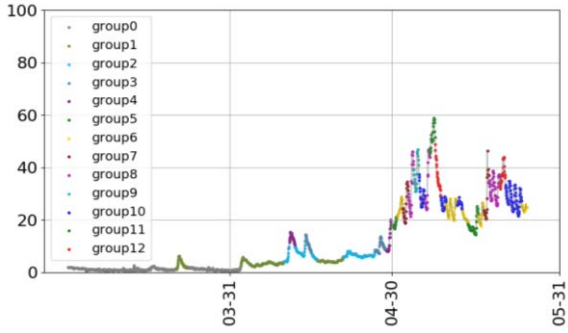

図-4 各グループの時系列グラフ（2010年）

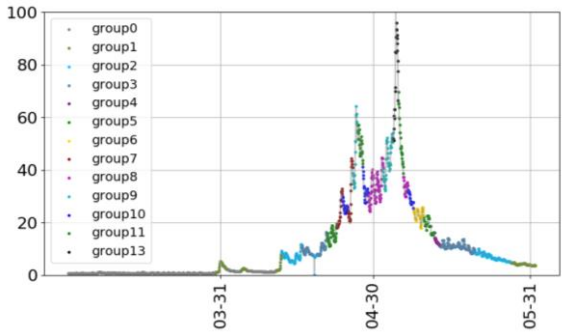

図-5 各グループの時系列グラフ（2012年）

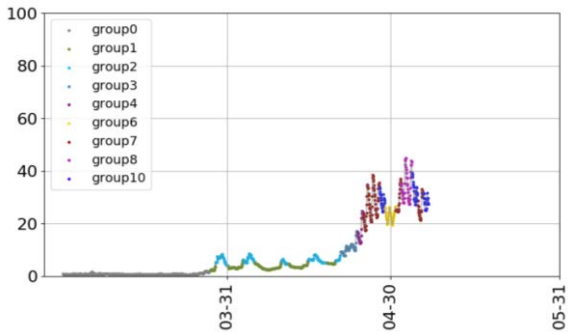

図-6 各グループの時系列グラフ（2014年）

表面温度，積雪重量，積雪深，融雪量の重要度が高く，日射量においては数日間のトレンドが重要であった．降水量においては，空間分布を司る埋め込み変数①，降水量の大小を司る埋め込み変数②ともに1日よりも1週間の変動の方がより重要度が高かった．これは，降水量の時空間分布は非線形性が強く，日変動のパターンではダム流入量のパターンをうまく説明できなかったことが考えられる．これは，重要度の計算にRFを用いていることに限界があるものと考えられる．逆に重要度の小さ

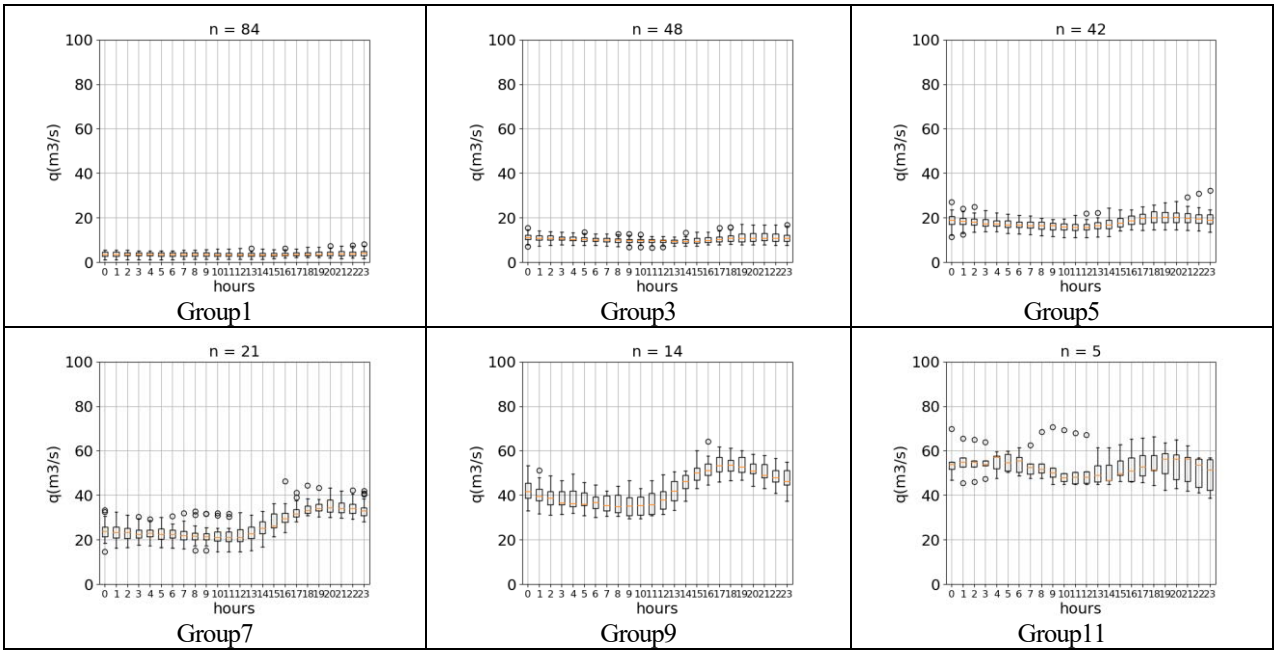

図-7　流入量の代表的な日パターンの分類結果（ボックスダイヤグラム）

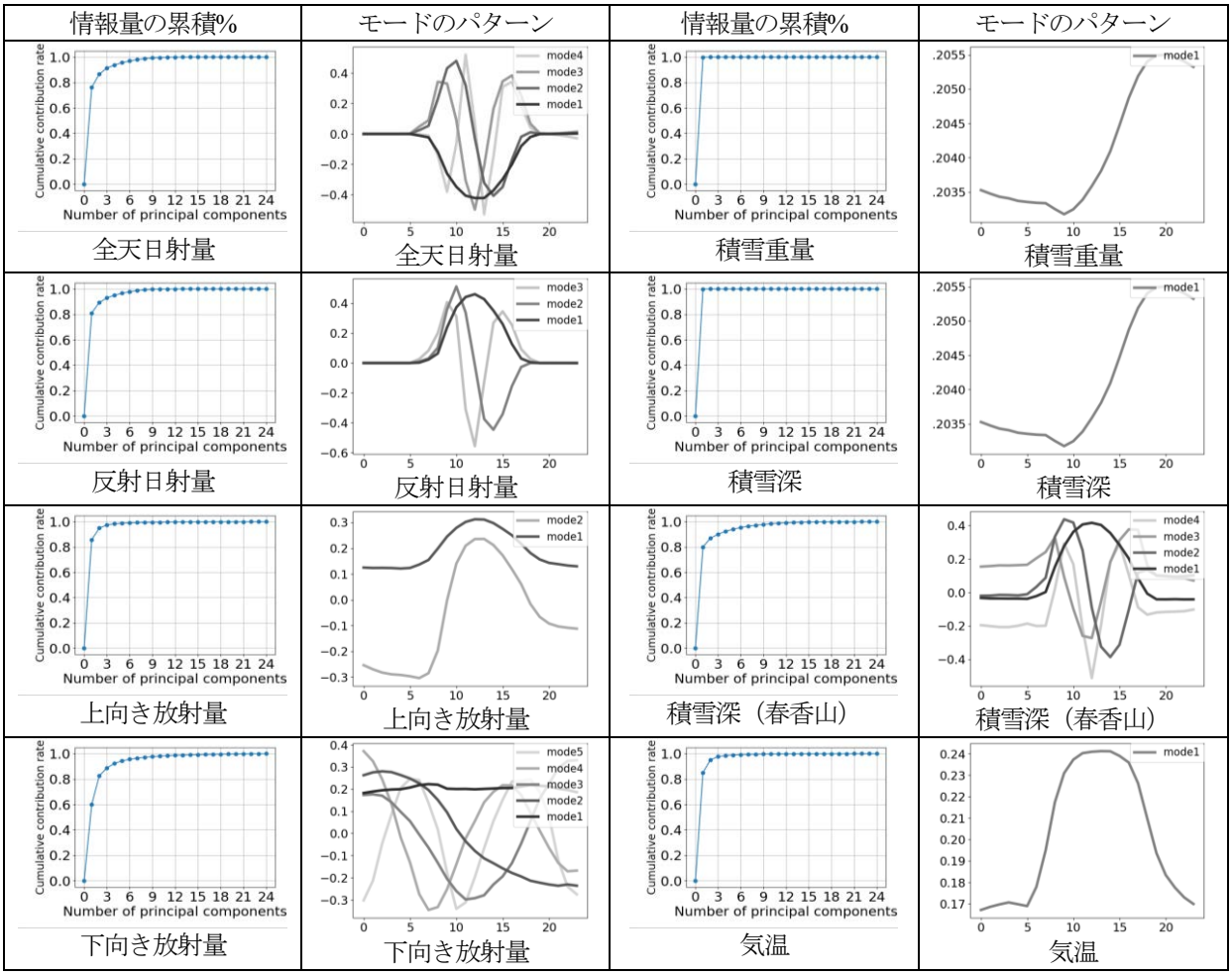

図-8　各項目の日変動のモードの可視化図

表-4　RFによる各変数の全モードの重要度

パターン抽出期間	項目	単位	モード数	重要度																		
				全モード平均	全モード最大	各モード																
						1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1日	全天日射量	W/m2	4	0.0050	0.0056	0.0	0.0	0.0	0.0														
	放射量(下向き)	W/m2	5	0.0055	0.0068	0.0	0.0	0.0	0.0	0.0													
	反射日射量	W/m2	3	0.0052	0.0077	0.0	0.0	0.0															
	放射量(上向き)	W/m2	2	0.0124	0.0195	0.0	0.0																
	表面温度	˚C	2	0.0150	0.0206	0.0	0.0																
	風速	m/s	12	0.0051	0.0068	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0						
	気温	˚C	1	0.0364	0.0364	0.0																	
	湿度	%	6	0.0049	0.0053	0.0	0.0	0.0	0.0	0.0	0.0												
	積雪重量(1,2平均)	kg/m2	1	0.0111	0.0111	0.0																	
	積雪深	cm	1	0.0299	0.0299	0.0																	
	融雪量	mm	4	0.0112	0.0159	0.0	0.0	0.0	0.0														
	積雪深	cm	1	0.0153	0.0153	0.0																	
	降水量空間分布（埋め込み①）	mm	18	0.0037	0.0051	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	降水量空間分布（埋め込み②）	mm	18	0.0037	0.0049	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1週間	全天日射量	W/m2	6	0.0068	0.0163	0.0	0.0	0.0	0.0	0.0	0.0												
	放射量(下向き)	W/m2	5	0.0059	0.0094	0.0	0.0	0.0	0.0	0.0													
	反射日射量	W/m2	6	0.0053	0.0097	0.0	0.0	0.0	0.0	0.0	0.0												
	放射量(上向き)	W/m2	2	0.0173	0.0283	0.0	0.0																
	表面温度	˚C	2	0.0220	0.0358	0.0	0.0																
	風速	m/s	6	0.0052	0.0082	0.0	0.0	0.0	0.0	0.0	0.0												
	気温	˚C	3	0.0225	0.0537	0.1	0.0	0.0															
	湿度	%	6	0.0052	0.0067	0.0	0.0	0.0	0.0	0.0	0.0												
	積雪重量(1,2平均)	kg/m2	1	0.0120	0.0120	0.0																	
	積雪深	cm	1	0.0123	0.0123	0.0																	
	融雪量	mm	5	0.0116	0.0232	0.0	0.0	0.0	0.0	0.0													
	積雪深	cm	1	0.0571	0.0571	0.1																	
	降水量空間分布（埋め込み①）	mm	6	0.0060	0.0115	0.0	0.0	0.0	0.0	0.0	0.0												
	降水量空間分布（埋め込み②）	mm	6	0.0058	0.0103	0.0	0.0	0.0	0.0	0.0	0.0												

表-5　各変数の重要度の最大値

項目	単位	パターン抽出期間	
		1日	1週間
全天日射量	W/m2	0.0056	0.0163
放射量(下向き)	W/m2	0.0068	0.0094
反射日射量	W/m2	0.0077	0.0097
放射量(上向き)	W/m2	0.0195	0.0283
表面温度	˚C	0.0206	0.0358
風速	m/s	0.0068	0.0082
気温	˚C	0.0364	0.0537
湿度	%	0.0053	0.0067
積雪重量(1,2平均)	kg/m2	0.0111	0.0120
積雪深	cm	0.0299	0.0123
融雪量	mm	0.0159	0.0232
積雪深	cm	0.0153	0.0571
降水量空間分布（埋め込み①）	mm	0.0051	0.0115
降水量空間分布（埋め込み②）	mm	0.0049	0.0103

表-6　重要度が0.01以上の抽出結果

項目	単位	パターン抽出期間	
		1日	1週間
全天日射量	W/m2		●
放射量(下向き)	W/m2		
反射日射量	W/m2		
放射量(上向き)	W/m2	○	●
表面温度	˚C	○	●
風速	m/s		
気温	˚C	○	●
湿度	%		
積雪重量(1,2平均)	kg/m2	○	●
積雪深	cm	●	○
融雪量	mm	○	●
積雪深（春香山）	cm	○	●
降水量空間分布（埋め込み①）	mm		●
降水量空間分布（埋め込み②）	mm		●

い情報は下向き放射量，風速，湿度があげられた．

これらの結果は概ね既往の知見から想定できるものであり，新規に意外な関係性を見出すことはできなかった．一方で，入力データの重要度を定量的に可視化できたことから，AIモデルの過学習防止に役立てていくことが可能であると考えられる．

５．まとめ

本研究では，もともと現象論が明らかとなっているデータに対して解析を行った．その結果としては，RFにより意外な関係性がみられるということではなく，熱収支法や流域の特徴などから想定できる結果であった．しかしながら，入力データの取捨選択が可能となり，AIモデルの過学習防止に役立つと考えられる．今後は，これらの結果に基づいて融雪期のダム流入量予測を行うAIモデルを構築する予定である．

参考文献

1) 滝口修司，キムスンミン，立川康人，市川温，萬和明：ニューラルネットワークを用いた積雪地域の河川流量予測における重要入力因子の抽出，土木学会論文集B1(水工学) Vol.74,No.4,pp877-882,2018.

2) Diederik P Kingma, Max Welling (2014): Auto-Encoding Variational Bayes, https://arxiv.org/abs/1312.6114 (preprint)

(2020. 4. 2受付)