نشریه پژوهش‌های حفاظت آب و خاک
جلد بیست و چهارم، شماره پنجم، ۱۳۹۶
http://jwsc.gau.ac.ir

مقایسه تطبیقی مدل‌های داده‌کاوی در ریزمقیاس‌نمایی بارش و دما (مطالعه موردی: حوضه آبخیز بازفت صمصامی)

نوید دهقانی[۱]، *هدی قاسمیه[۲]، سیدجواد ساداتی‌نژاد[۳]، خلیل قربانی[۴] و علی‌اصغر بسالت‌پور[۵]

[۱]دانشجوی دکتری گروه مرتع و آبخیزداری، دانشگاه کاشان، [۲]استادیار گروه مرتع و آبخیزداری، دانشگاه کاشان،
[۳]دانشیار گروه انرژی‌های نو و محیط زیست، دانشگاه تهران، [۴]دانشیار گروه مهندسی آب، دانشگاه علوم کشاورزی و منابع طبیعی گرگان،
[۵]استادیار گروه علوم خاک، دانشگاه ولی‌عصر (عج) رفسنجان

تاریخ دریافت: ۹۵/۸/۲۵ ؛ تاریخ پذیرش: ۹۶/۱۰/۹

چکیده

سابقه و هدف: دما و بارش به‌عنوان دو متغیر مهم هواشناسی، به‌خصوص در مناطق خشک و نیمه‌خشک مطرح هستند. در نتیجه، تعیین مقدار این متغیرها، تغییرات آن‌ها و پیش‌بینی این پدیده‌ها به‌منظور برنامه‌ریزی دقیق‌تر در مدیریت بخش‌های کشاورزی، اقتصادی و اجتماعی، ضروری می‌باشد. امروزه عدم تطابق مقیاس مکانی و زمانی مورد نیاز در مدل‌های بررسی‌کننده تأثیر تغییر اقلیم با خروجی مدل‌های GCM و نیاز به بررسی روند تغییر در متغیرهای حدی هواشناسی در مقیاس منطقه‌ای، باعث شده است تا روش‌های ریزمقیاس‌نمایی مختلفی توسعه یابند. از این‌رو هدف از این پژوهش، مقایسه تطبیقی مدل‌های داده‌کاوی در ریزمقیاس‌نمایی بارش و دما براساس داده‌های مدل گردش عمومی NCEP است.

مواد و روش‌ها: منطقه مورد مطالعه در این پژوهش، حوضه آبخیز بازفت صمصامی است. این حوضه، یکی از زیرحوضه‌های کارون شمالی است که در شمال‌غربی استان چهارمحال و بختیاری واقع شده است. ایستگاه‌های باران‌سنجی و هیدرومتری مرغک در خروجی آن واقع شده است. در این پژوهش، کارایی چهار روش درخت تصمیم (M5)، نزدیک‌ترین همسایه (KNN)، روش پرسپترون چندلایه (MLP) و رگرسیون خطی ساده (SLR) برای مدل‌سازی بارش و دمای ماهانه ایستگاه مرغک در دوره آموزش ۱۹۹۰-۱۹۷۱ و دوره آزمون ۱۹۹۱-۲۰۰۰ با استفاده از پارامترهای خروجی NCEP مورد ارزیابی قرار گرفت.

یافته‌ها: نتایج مدل‌سازی بارش ماهانه با استفاده از مدل‌های مذکور نشان داد که خروجی همه مدل‌ها به‌جز مدل KNN، مقادیر منفی را برای بارش ارائه می‌کنند. پیش‌بینی بارش توسط مدل درخت تصمیم در ماه‌های میلادی ژانویه، مارس، آوریل و دسامبر، دارای میانگین کم‌تری نسبت به مقادیر مشاهده شده (P) است. این وضعیت در سایر مدل‌ها نیز تا حدودی مشاهده می‌شود. همچنین با توجه به این‌که حد پایین بارش صفر است، از کم بودن مقادیر پیش‌بینی‌شده نسبت به مقادیر مشاهده شده می‌توان نتیجه گرفت که مقادیر حدی بیشینه بارش با این مدل‌ها به‌خوبی

* مسئول مکاتبه: h.ghasemieh@kashanu.ac.ir

پیش‌بینی نشده است. پیش‌بینی بارش توسط همه مدل‌ها در همه ماه‌ها به‌جز ماه مه، دارای انحراف معیار کم‌تری نسبت به مقادیر مشاهده‌شده (P) است. نتایج پیش‌بینی دمای ماهانه نیز نشان داد که تنها خروجی MLP، مقادیر منفی را برای دمای ماهانه ارائه می‌کند که این می‌تواند به‌دلیل خاصیت برون‌یابی و تعمیم در روش پرسپترون چندلایه باشد. همچنین انحراف معیار به‌دست آمده از تمامی مدل‌ها در ماه‌های ژانویه، فوریه، مارس، آوریل، ژوئیه، اوت، اکتبر، نوامبر و دسامبر بیش‌تر از انحراف معیار دمای مشاهده شده است. نتایج تحلیل‌های آماری نیز نشان داد که مدل درخت تصمیم در مرحله آزمون با توجه به معیارهای ریشه میانگین مربعات خطا، میانگین خطای اریب و ضریب همبستگی نسبت به مدل‌های دیگر، برآورد بهتری برای بارش و دمای ماهانه داشته است. اگرچه نتایج ضریب تعیین این مدل در مرحله آزمون برای برآورد دمای ماهانه، ضعیف‌تر از بارش ماهانه می‌باشد.

نتیجه‌گیری: نتایج بررسی کارایی چهار مدل KNN، M5، SLR و MLP در مدل‌سازی بارش و دمای ماهانه ایستگاه هواشناسی مرغک با داده‌های خروجی مدل NCEP، بیانگر ضعف این مدل‌ها در ریزمقیاس‌نمایی بارش و دمای ماهانه بود. بنابراین با وجود برتری نسبی مدل درخت تصمیم M5 نسبت به سایر مدل‌ها، استفاده از مدل‌های داده‌کاوی مذکور برای پیش‌بینی بارش و دما در ایستگاه مرغک توصیه نمی‌شود.

واژه‌های کلیدی: ریزمقیاس‌نمایی، درخت تصمیم (M5)، نزدیک‌ترین همسایه (KNN)، روش پرسپترون چندلایه (MLP)، رگرسیون خطی ساده (SLR)

مقدمه

صنعتی شدن جوامع و افزایش گازهای گل‌خانه‌ای در دهه‌های گذشته، باعث افزایش دمای کره زمین و تغییر در سایر پارامترهای اقلیمی شده است که در نوشته‌های علمی، به آن پدیده تغییر اقلیم[1] اطلاق می‌شود. طبق گزارش هیأت بین‌الدول تغییر اقلیم[2]، طی دوره صد ساله منتهی به سال ۲۰۰۵، دمای متوسط جهانی به‌میزان ۰/۷٤ درجه سانتی‌گراد افزایش یافته است (۹). افزایش دمای سطح زمین و تغییرات در الگوهای بارندگی، پدیده‌های قالب در تغییر اقلیم می‌باشد که این دو، تقریباً تمام بخش‌های دیگر چرخه آب را تحت‌تأثیر قرار می‌دهد. تمام مدل‌های سه‌بعدی جفت‌شده جوی- اقیانوسی گردش عمومی هوا (AOGCM[3])، افزایش دما در سطح زمین و افزایش شدت بارش و مقدار آن بر اثر افزایش غلظت گازهای گل‌خانه‌ای را در قرن حاضر پیش‌بینی می‌کنند (۸). روش‌های مختلفی برای شبیه‌سازی متغیرهای اقلیمی در دوره‌های آتی وجود دارد که معتبرترین آن‌ها، استفاده از خروجی مدل‌های اتمسفر- اقیانوس گردش عمومی جو می‌باشد. عدم تطابق مقیاس مکانی و زمانی مورد نیاز در مدل‌های بررسی‌کننده تأثیر تغییر اقلیم با خروجی مدل‌های GCM و نیاز به بررسی روند تغییر در پارامترهای حدی هواشناسی در مقیاس منطقه‌ای، باعث شده است تا روش‌های ریزمقیاس‌نمایی[4] مختلفی توسعه یابند (۹).

ریزمقیاس‌نمایی، فرآیند انتقال اطلاعات اقلیمی از یک مدل اقلیمی درشت‌مقیاس به ریزمقیاس است (۱۳). به‌عبارتی این مدل‌ها با استفاده از خروجی مدل‌های گردش عمومی و به‌کارگیری سناریوی خاص مدل تولیدکننده داده‌های آب و هوایی، داده‌های

1- Climate Change
2- IPCC: Intergovernmental Panel on Climate Change
3- AOGCM: Atmospheric-Ocean General Circulation Model
4- Downscaling

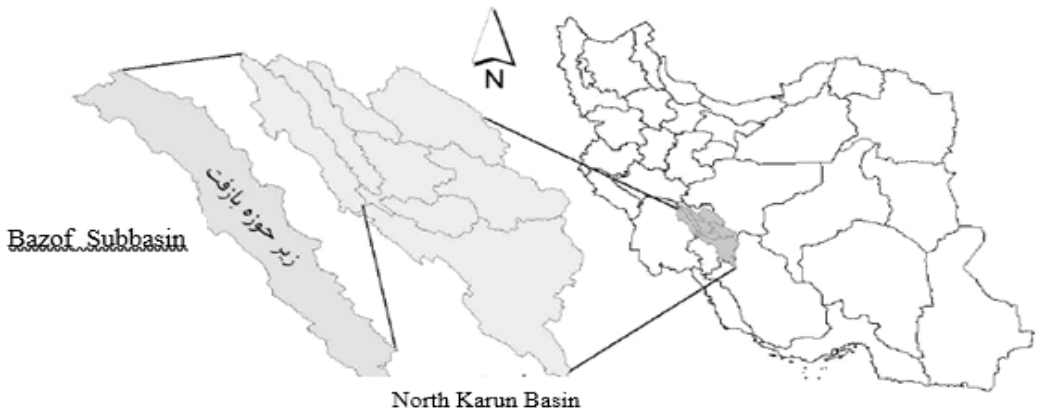
گردش عمومی در مقیاس درشت را به مقیاس‌های ریزتر تبدیل می‌کنند. مهم‌ترین نقاط قوت این مدل‌ها، ارزان بودن، سرعت بالا و امکان استفاده از آن‌ها بدون نیاز به ابررایانه‌ها و یا رایانه‌های بسیار سریع می‌باشند (۱۲). از معروف‌ترین مدل‌های آماری که برای ریزمقیاس‌نمایی داده‌های اقلیمی استفاده می‌شود، مدل LARS_WG[1] و SDSM[2] هستند که به‌صورت بسته‌های نرم‌افزاری و رایگان در اختیار هستند. قابلیت این مدل‌ها در ریزمقیاس‌نمایی متغیرهای دما و بارش توسط پژوهشگران مختلف ارزیابی شده است (۱۵). به‌طور نمونه دیبایک و کولیبالی (۲۰۰۶)، دو روش شبکه‌های عصبی مصنوعی و SDSM را به‌منظور ریزمقیاس‌کردن داده‌های بارش و درجه حرارت روزانه در حوضه ساگونوی در شمال ایالت کبک کانادا مقایسه کردند و به این نتیجه رسیدند که شبکه عصبی مصنوعی در ریزمقیاس‌نمایی داده‌های بیشینه و کمینه دمای روزانه و همچنین بارش روزانه نسبت به روش SDSM، پیش‌بینی بهتری دارد (۵). تریپاتی و همکاران (۲۰۰۶) با مقایسه روش‌های شبکه عصبی مصنوعی و ماشین بردار پشتیبان[3] برای ریزمقیاس‌نمایی بارش ماهانه در هندوستان، روش ماشین بردار پشتیبان را مناسب دانستند (۱۷). خان و همکاران (۲۰۰۶) سه روش ریزمقیاس‌نمایی شامل مدل آماری SDSM، مدل تولید داده LARS-WG و شبکه عصبی مصنوعی را برای ریزمقیاس‌نمایی متغیرهای بارش روزانه و دماهای حداقل و حداکثر روزانه به‌کار بردند. ایشان از داده‌های ۴۰ ساله مشاهده شده در زیرحوضه چوت‌دو دایابل کانادا و داده‌های مدل گردش عمومی NCEP[4] در دوره آماری ۲۰۰۱-۱۹۶۱ برای مدل‌سازی استفاده کردند. سرانجام نتایج پژوهش ایشان نشان داد که مدل SDSM، بهترین نتایج را در بازسازی خصوصیات داده‌های مشاهده شده، ارائه کرده و مدل شبکه عصبی مصنوعی از این نظر، کم‌ترین کارایی را داشته است (۱۰). آکسورن سینگچای و سرینیلتا (۲۰۱۱)، سه روش ریزمقیاس‌نمایی آماری را برای پیش‌بینی دما و بارش در ۴۵ ایستگاه هواشناسی در تایلند در دوره آماری ۲۰۰۷-۱۹۶۵ مقایسه کردند. ایشان از داده‌های دوباره آنالیز شده آزمایشگاه ژئوفیزیکی دینامیک سیال[5] استفاده کردند و نتیجه گرفتند که روش ماشین بردار پشتیبان با تابع کرنل پایه شعاعی نسبت به روش‌های ماشین بردار پشتیبان با تابع کرنل چندجمله‌ای و روش رگرسیون چندمتغیره، نتیجه بهتری را ارائه می‌کند (۱). دیپاشری و موجومدار (۲۰۱۱) نیز سه روش میدان تصادفی مشروط (CRF)[6]، نزدیک‌ترین همسایگی K (KNN)[7] و ماشین بردار پشتیبان (SVM) را برای ریزمقیاس‌نمایی بارش روزانه به کمک داده‌های مدل گردش عمومی NCEP در دوره آماری ۲۰۰۴-۱۹۵۱ در ایالت پنجاب مورد مقایسه قرار دادند و نتیجه گرفتند که مدل‌های CRF و KNN در مقایسه با مدل SVM از لحاظ آماری، نتایج بهتری را ارائه می‌کند (۴). چن و همکاران (۲۰۱۲)، روش ماشین بردار پشتیبان و شبکه عصبی مصنوعی را در ریزمقیاس‌نمایی داده‌های بارش در رودخانه یانگ‌تسه در چین مورد ارزیابی قرار دادند و نتیجه گرفتند که ماشین بردار پشتیبان با تابع کرنل پایه شعاعی، روش بهتری برای ارزیابی اثرات تغییر اقلیم در هیدرولوژی است (۲). در پژوهشی دیگر قمقامی (۱۳۸۹)، از یک روش ناپارامتری مبتنی بر برآوردگر کرنل برای شبیه‌سازی بارش ماهانه استفاده کرد. نتایج پژوهش وی نشان داد که این روش، قابلیت شبیه‌سازی مناسب میانگین‌ها و واریانس‌های ماهانه و همچنین وقایع حدی بارش مانند بارندگی‌های با دوره

1- LARS-WG: A Stochastic Weather Generator for Climate Change Impact Assessments
2- SDSM: Statistical Downscaling *Model*
3- SVM: Support Vector Machines
4- NCEP: National Centers for Environmental Prediction
5- GFDL: Geophysical Fluid Dynamic Laboratory
6- CRF: Conditional Random Field
7- KNN: K-Nearest Neighbor

بازگشت بالا را دارد (٦). قربانی (۱۳۹۳)، کارایی سه روش ماشین بردار پشتیبان (SVM)، درخت تصمیم (M5) و نزدیک‌ترین همسایه K (KNN) را در یزمقیاس‌نمایی بارش در ایستگاه کرمانشاه بررسی کردند. نتایج این بررسی نشان داد که بارش شبیه‌سازی‌شده با هر یک از مدل‌های داده‌کاوی، دارای میانگین و انحراف معیار کم‌تری نسبت به داده‌های مشاهداتی هستند و مقادیر حدی را نمی‌توانند به خوبی پیش‌بینی کنند. با این وجود روش نزدیک‌ترین همسایگی K نسبت به دیگر روش‌ها، نتایج بهتری را ارائه کرد (۷). با توجه به مطالب ارائه شده، هدف از این پژوهش، بررسی چهار مدل درخت تصمیم (M5)، نزدیک‌ترین همسایه K (KNN)، روش پرسپترون چندلایه (MLP) و روش رگرسیون خطی ساده (SLR) در ایستگاه مرغک حوضه آبخیز بازفت صمصامی از زیرحوضه‌های کارون شمالی است تا دقت این مدل‌ها برای ریزمقیاس‌نمایی دو متغیر مهم برای مطالعات اقلیمی و هیدرولوژیکی مورد بررسی قرار گیرد.

مواد و روش‌ها

منطقه مورد مطالعه: حوضه آبخیز بازفت صمصامی، یکی از زیرحوضه‌های کارون شمالی است که در شمال‌غربی استان چهارمحال و بختیاری واقع شده است. وسعت این حوضه، ۲۱۳۳ کیلومترمربع؛ بارش متوسط سالانه، ۹٦٦ میلی‌متر است. ایستگاه باران‌سنجی و هیدرومتری مرغک در خروجی حوضه مطالعاتی واقع شده است. شکل ۱، موقعیت منطقه مطالعاتی در حوضه آبخیز کارون شمالی را نشان می‌دهد. در این پژوهش داده‌های بارش و دمای متوسط روزانه ایستگاه مرغک طی سال‌های ۲۰۰۰-۱۹۷۱ بر طبق داده‌های موجود مدل گردش عمومی جو NCEP از پایگاه CCIS گزارش چهارم IPCC مشتمل بر ۲٦ متغیر تهیه گردید. سپس انتخاب پیش‌بینی‌کننده‌های مؤثر با استفاده از مدل SDSM انجام شد. به این صورت که در این مدل با روش سعی و خطا، بالاترین همبستگی و کم‌ترین واریانس خطا بین متغیرهای پیش‌بینی‌کننده و متغیرهای پیش‌بینی‌شونده محاسبه می‌شود تا مؤثرترین پیش‌بینی‌کننده‌ها انتخاب گردند. در جدول ۱، بهترین متغیرهای پیش‌بینی‌کننده برای دما و بارش در این منطقه که دارای بالاترین همبستگی بر روی شبکه مدل HadCM3 هستند، نشان داده می‌شود.



شکل ۱- موقعیت جغرافیایی حوضه آبخیز بازفت صمصامی در حوضه کارون شمالی.

Figure 1. Geographical location of Bazoft-e-Samsami watershed in Northern Karun basin.

جدول ۱- متغیرهای پیش‌بینی‌کننده منتخب دما و بارش.

Table 1. Selected predictor variables of temperature and rainfall.

	نام (Name)	ویژگی (characteristic)
بهترین متغیرهای پیش‌بینی‌کننده دما (The best predictor variables of temperature)	ncepp_zna	حالت گردابی سطحی (زمان/۱) Surface vorticity (S^{-1})
	ncepp5_zn	حالت گردابی ۵۰۰ hpa (زمان/۱) 500 hPa vorticity (S^{-1})
	nceprhumna	رطوبت نسبی سطحی (کیلوگرم بر کیلوگرم) Surface relative humidity (kg/kg)
	ncepsphuna	رطوبت مشخص سطحی (کیلوگرم بر کیلوگرم) Surface specific humidity (kg/kg)
بهترین متغیرهای پیش‌بینی‌کننده بارش (The best predictor variables of rainfall)	ncepmslpna	متوسط فشار سطح دریا (پاسگال) Mean sea level pressure (Pa)
	ncepp5_una	سرعت ناحیه‌ای 500hPa (متر بر ثانیه) 500 hPa zonal speed (m/s)
	ncepp5_vna	سرعت نصف‌النهاری 500hPa (متر بر ثانیه) 500 hPa meridional speed (m/s)
	ncepp500na	رطوبت نسبی 500hPa (کیلوگرم بر کیلوگرم) 500 hPa relative humidity (kg/kg)

ریزمقیاس‌نمایی با استفاده از مدل‌های داده‌کاوی: داده‌کاوی، فرآیندی است که ابزارهای مختلف تحلیل داده را به‌کار می‌گیرد تا الگوها و رابطه‌های فیزیکی متغیرها را در مجموعه داده‌های مختلف کشف کند (۱۸). الگوریتم‌های زیادی در داده‌کاوی برای هدف‌های مختلف مانند طبقه‌بندی و پیش‌بینی استفاده می‌شود. در این پژوهش، چهار روش مدل درخت تصمیم M5، رگرسیون خطی ساده (SLR)، نزدیک‌ترین همسایگی K و روش پرسپترون چندلایه (MLP) مورد استفاده و ارزیابی قرار گرفتند.

مدل درخت تصمیم M5[1]: درخت‌های تصمیم، روشی برای نمایش یک سری از قانون‌ها هستند که منتهی به یک رده یا مقدار می‌شوند. درخت‌های تصمیم به کمک جداسازی متوالی داده‌ها، به یک سری گروه جداگانه تبدیل شده و سعی می‌شود فاصله بین گروه‌ها در فرآیند جداسازی، افزایش یابد. ساختار یک مدل درختی، شامل ریشه، گره‌های داخلی و برگ است. از مدل‌های درخت تصمیم در حل بسیاری از مسائل طبقه‌بندی و رگرسیون استفاده شده است. برای اولین بار کوینلان (۱٤)، مدل درخت تصمیم موسوم به M5 را برای پیش‌بینی داده‌های پیوسته ارائه کرد. این مدل، بر خلاف مدل‌های درخت تصمیم معمول که کلاس یا رده‌های گسسته را به‌عنوان خروجی ارائه می‌کنند، یک مدل خطی چندمتغیره را برای داده‌ها در

1- M5 Rules

هر گره از مدل درختی می‌سازد. تشکیل ساختار مدل‌های درخت تصمیم‌گیری، شامل مراحل ایجاد درخت و هرس کردن آن است (۲۰). در مدل M5 از یک جستجوی حریصانه برای حذف متغیرهایی که مشارکت کمی در مدل دارند، استفاده می‌شود. البته گاهی وقت‌ها همه متغیرها حذف شده و فقط یک مقدار ثابت باقی می‌ماند (۱۳).

نزدیک‌ترین همسایگی K (KNN): یکی دیگر از روش‌های مدل‌سازی در داده‌کاوی، الگوریتم نزدیک‌ترین همسایگی K است. این الگوریتم، جزء روش‌های یادگیری نظارت شده است که هم در طبقه‌بندی و هم در پیش‌بینی استفاده می‌شود. نحوه عملکرد این الگوریتم، براساس مشاهدات و نمونه‌ها می‌باشد. براساس این الگوریتم می‌توان یک نمونه جدید را براساس اکثریت K گروه و دسته که نزدیک‌ترین همسایگی‌ها را با آن نمونه داشته باشند، تقسیم‌بندی کرد. به‌عبارت دیگر می‌توان گفت که این روش، K تعداد از الگوهای مشابه را پیدا کرده و براساس آن‌ها، ارزش نمونه مورد مطالعه را پیش‌بینی می‌کند. این الگوریتم، جزء روش‌های تنبل به حساب می‌آید. به این دلیل که مرحله آموزش را همان زمان که نمونه جدید باید طبقه‌بندی شود، اجرا می‌کند. بر این اساس، KNN نسبت به دیگر الگوریتم‌های یادگیری محاسبات بیش‌تری را لازم دارد. KNN برای داده‌های پویا و داده‌هایی که سریع تغییر می‌کنند و به روز می‌شوند، مناسب است (۱۹).

رگرسیون خطی ساده (SLR)[1]: رگرسیون، یکی از مهم‌ترین مباحث آماری در جغرافیا و مطالعات اقلیمی است و یکی از اهداف آن در اقلیم‌شناسی، انتخاب بهترین مدل مناسب برای یافتن روابط موجود بین داده‌ها و تهیه مدل‌هایی برای اهداف پیش‌بینی است. در این مدل آماری، فرض بر این است که رابطه بین متغیرهای مستقل و متغیر وابسته، به‌صورت زیر است (۱۱):

$$Y_i=ß_o+ß_1x+e \qquad (۲)$$

که در آن، $ß_o$ عرض از مبدأ، ß شیب خط، Y متغیر وابسته و e عامل خطا می‌باشد. رگرسیون خطی، به دو صورت رگرسیون خطی ساده و رگرسیون خطی چندمتغیره یا چندگانه طرح می‌گردد. رگرسیون خطی ساده، به پیش‌بینی مقدار یک متغیر وابسته براساس مقدار یک متغیر مستقل می‌پردازد؛ اما رگرسیون چندگانه، روشی است برای تحلیل مشارکت دو یا چند متغیر مستقل در تغییرات یک متغیر وابسته (۱۱).

شبکه عصبی پرسپترون چندلایه (MLP)[2]: این شبکه، رایج‌ترین شبکه‌های عصبی و جزء شبکه‌های عصبی پیشخور می‌باشند که قادرند با انتخاب مناسب تعداد لایه‌ها و نرون‌ها، یک نگاشت غیرخطی را با دقت دلخواه انجام دهند. پارامترهای قابل تنظیم در شبکه‌های MLP، وزن اتصالات مابین لایه‌هامی‌باشد و فرآیند آموزش در این شبکه‌ها، به معنی یافتن مقادیر مناسب برای وزن‌های اتصالات مابین نرون‌ها است. متداول‌ترین الگوریتم یادگیری این شبکه‌ها، الگوریتم پس‌انتشار خطا است. از مهم‌ترین قسمت‌های تعیین ساختار بهینه شبکه پرسپترون چند لایه، تعیین تعداد لایه‌های پنهان و تعداد نرون‌های هر لایه پنهان برای دست‌یابی به کم‌ترین خطا می‌باشد. قضیه‌ای در تئوری‌های مربوط به شبکه عصبی مصنوعی وجود دارد که اثبات می‌کند که یک لایه پنهان با تعداد نرون کافی، قادر به تخمین هر رابطه غیرخطی است. بین نرون‌های لایه‌های مختلف، اتصالاتی وجود دارد که

1- SLR: Simple Linear Regression

2- MLP: Multilayer Perceptron Network

هر کدام، دارای وزن‌هایی می‌باشند. طی فرایند آموزش، این وزن‌ها و مقادیر ثابتی که با آن‌ها جمع می‌شود و در اصطلاح بایاس نامیده می‌شود، به‌طور پی در پی تغییر می‌کنند تا خطای بین مقادیر تخمین زده شده و مقادیر واقعی به حداقل مقدار خود برسد. برای انتقال خروجی‌های هر لایه به لایه‌های بعدی، از توابع محرک استفاده می‌شود. توابع محرک، انواع مختلفی دارند که از معروف‌ترین آن‌ها می‌توان به تابع خطی، تابع سیگموئید، تابع تانژانت هایپربولیک و ... اشاره کرد (۳).

ارزیابی مدل‌ها: بعد از مرتب‌سازی داده‌ها و تشکیل جدول اطلاعاتی، متغیرهای بارش و دمای متوسط ماهانه در ایستگاه مرغک به‌عنوان متغیر وابسته و داده‌های مدل گردش عمومی NCEP به‌عنوان متغیر مستقل به مدل‌های داده‌کاوی به‌صورت مجزا معرفی شدند. دوره آماری ۱۹۹۰-۱۹۷۱، به‌عنوان دوره آموزش و دوره آماری ۲۰۰۰-۱۹۹۱ به‌عنوان دوره آزمون در نظر گرفته شدند.

با در نظر گرفتن مقادیر مشاهده شده ($Z(x_i)$) و مقادیر پیش‌بینی شده ($\hat{Z}(x_i)$) توسط هر یک از مدل‌ها بعد از اجرای آن‌ها، از دو معیار ریشه میانگین مربعات خطا (RMSE)[1] و میانگین خطای اریبی (MBE)[2] و ضریب تعیین (R^2)[3] برای ارزیابی مدل‌ها استفاده شد. معیار RMSE بزرگی خطا و معیار MBE، میزان انحراف از خط نیمساز را نشان می‌دهد (۷).

$$RMSE = \left(\sum_{i=1}^{n} (\hat{Z}(x_i) - Z(x_i))^2 / n \right)^{1/2} \quad (۳)$$

$$MBE = \sum_{i=1}^{n} (\hat{Z}(x_i) - Z(x_i)) / n \quad (٤)$$

$$R^2 = \frac{\left[\sum_{i=1}^{n} (Z_{x_i} - \overline{Z_{xi}})(\hat{Z}_{x_i} - \overline{Z_{xi}}) \right]^2}{\sum_{i=1}^{n} (Z_{x_i} - \overline{Z_{xi}})^2 \cdot \sum_{i=1}^{n} (\hat{Z}_{x_i} - \overline{Z_{xi}})^2} \quad (۵)$$

نتایج و بحث

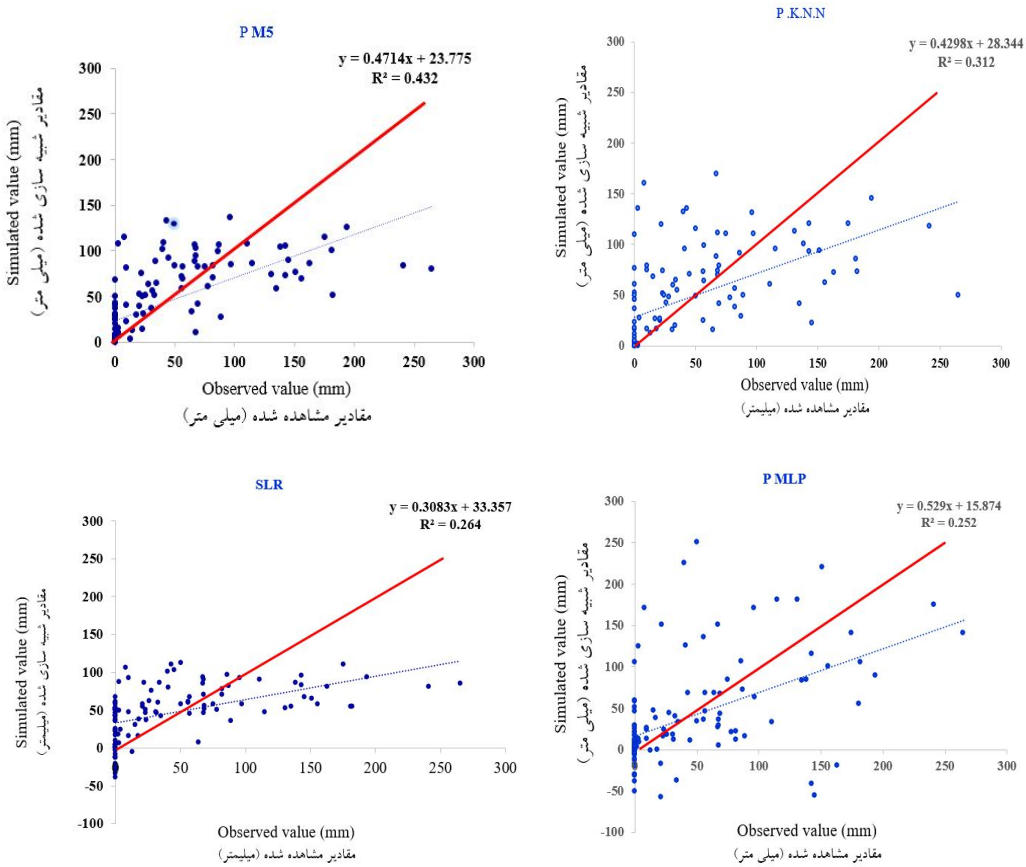
در مدل‌سازی داده‌ها، ابتدا مقادیر پیش‌بینی‌شده بارش ماهانه با هر یک از مدل‌های مورد مطالعه در مقابل مقادیر مشاهده ترسیم شدند (شکل ۲). از خط نیمساز نیز (خط قرمزرنگ) برای میزان انحراف مقادیر پیش‌بینی‌شده از مقادیر مشاهده‌شده استفاده شد. نتایج به‌دست آمده نشان داد که مدل‌سازی با مدل درخت تصمیم با توجه به خط برازش داده شده (خط سیاه‌رنگ) و ضریب R^2 نمایش داده شده روی نمودارها (R^2=۰/٤۳۲)، بهتر از نتایج سایر مدل‌ها به‌دست آمده است. همچنین خروجی همه مدل‌ها به‌جز مدل KNN، مقادیر منفی را نیز برای بارش ارائه می‌کنند که این موضوع می‌تواند به‌دلیل خاصیت برون‌یابی و تعمیم در رگرسیون باشد که در روش‌های درخت تصمیم، رگرسیون خطی ساده و پرسپترون چندلایه استفاده می‌شود. همچنین مدل درخت تصمیم (M5)، حداکثر تا حدود ۷۹/۷۷ میلی‌متر، روش نزدیک‌ترین همسایه (KNN) تا حدود ۵۰/۱ میلی‌متر، روش رگرسیون خطی ساده (SLR) تا حدود ۸٤/٥٦ میلی‌متر و در نهایت روش پرسپترون چندلایه (MLP) تا حدود ۱٤۰/۹۱ میلی‌متر را برای بارش ماهانه پیش‌بینی می‌کنند، در حالی‌که تا ۲٦٥ میلی‌متر بارش ماهانه در ایستگاه مورد نظر گزارش شده است (شکل ۲). این بخش از نتایج به‌دست آمده با نتایج

1- RMSE: Root Mean Square Error
2- MBE: Mean Bias Error
3- Determination Coefficient

قربانی (۱۳۹۳)، مطابقت دارد. میانگین و انحراف معیار ماهانه بارش نیز به تفکیک ماه‌های سال ترسیم شدند که در شکل‌های ۳ و ٤ آورده شده است. نتایج نشان داد که تمامی مدل‌های مورد مطالعه در این پژوهش، در پیش‌بینی بارندگی در ماه‌های پربارش موفق عمل نکرده‌اند. همچنین نتایج نشان داد که پیش‌بینی بارش توسط مدل درخت تصمیم در ماه‌های میلادی ژانویه، مارس، آوریل و دسامبر، دارای میانگین کم‌تری نسبت به مقادیر مشاهده شده (P) است. این وضعیت در سایر مدل‌ها نیز تا حدودی مشاهده می‌شود. همچنین با توجه به این‌که حد پایین بارش صفر است، از کم بودن مقادیر پیش‌بینی‌شده نسبت به مقادیر مشاهده‌شده می‌توان نتیجه گرفت که مقادیر حدی بیشینه بارش با این مدل‌ها به‌خوبی پیش‌بینی نشده است (شکل ٤). همچنین نتایج نشان داد که پیش‌بینی بارش توسط همه مدل‌ها در همه ماه‌ها به‌جز ماه مه دارای انحراف معیار کم‌تری نسبت به مقادیر مشاهده شده (P) است. نتایج این بخش از پژوهش، تقریباً با نتایج قربانی (۱۳۹۳)، مدرسی (۱۳۸۸) و نیز نتایج آکسوون سینگچای و سرینیلتا (۲۰۱۱) مطابقت دارد.

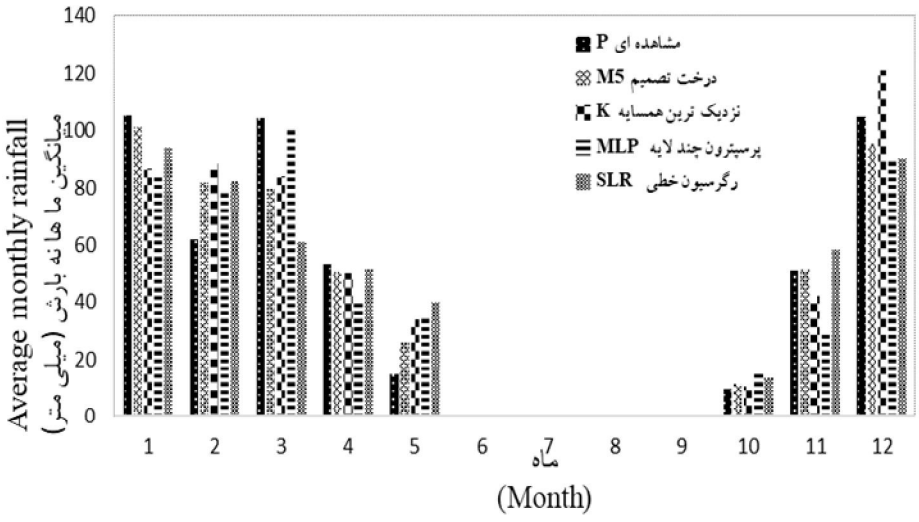
همچنین نتایج به‌دست آمده از شکل ٤ نشان می‌دهد که مدل درخت تصمیم نسبت به سایر مدل‌ها توانسته است در ماه‌های ژانویه، فوریه، آوریل، مه، اکتبر، نوامبر و دسامبر، مقدار بارش ماهانه را با اختلاف کم‌تری نسبت به مقادیر بارش مشاهده شده پیش‌بینی کند و نتایج در سایر ماه‌ها، کمی فرق کرده است. پس از مدل‌سازی بارش، نوبت به مدل‌سازی ماهانه دما رسید. در این بخش نیز، ابتدا مقادیر پیش‌بینی‌شده دمای ماهانه با هر یک از مدل‌های مورد مطالعه در مقابل مقادیر مشاهده ترسیم شدند (شکل ۵) و از خط نیمساز (خط قرمزرنگ) نیز برای بررسی میزان انحراف مقادیر پیش‌بینی‌شده از مقادیر مشاهده‌شده استفاده شد. نتایج نشان داد که تمامی مدل‌های مورد مطالعه در این پژوهش، در پیش‌بینی داده‌های دما موفق عمل نکرده‌اند و حتی نتایج پیش‌بینی دمای ماهانه با مدل‌ها، ضعیف‌تر از نتایج پیش‌بینی بارش ماهانه با استفاده از مدل‌های مذکور است.

البته در مورد دما نیز مانند بارش، نتایج مدل درخت تصمیم با توجه به خط برازش داده شده و ضریب R^2 نمایش داده شده روی نمودارها (R^2=۰/۱۹۸)، بهتر از نتایج سایر مدل‌ها به‌دست آمده است. تمامی روش‌ها به‌جز روش پرسپترون چندلایه (MLP) برای پیش‌بینی دمای ماهیانه، مقادیر مثبت دارند و برآورد منفی نداشته‌اند. تنها خروجی MLP، مقادیر منفی را برای دمای ماهانه ارائه می‌کند که این می‌تواند به‌دلیل خاصیت برون‌یابی و تعمیم در روش پرسپترون چندلایه باشد. همچنین مدل درخت تصمیم (M5)، حداکثر تا حدود ۱۸/۵ درجه سانتی‌گراد، روش نزدیک‌ترین همسایه K (KNN) تا حدود ۲۲/۳۲ درجه سانتی‌گراد، روش رگرسیون خطی ساده (SLR) تا حدود ۱۸/۸۵ درجه سانتی‌گراد و در نهایت روش پرسپترون چندلایه (MLP) تا حدود ۹/۲۷ درجه سانتی‌گراد را برای دمای ماهانه پیش‌بینی می‌کنند، در حالی‌که تا حدود ۳٤/۸٦ درجه سانتی‌گراد دمای ماهانه در ایستگاه مورد نظر گزارش شده است (شکل ۵).



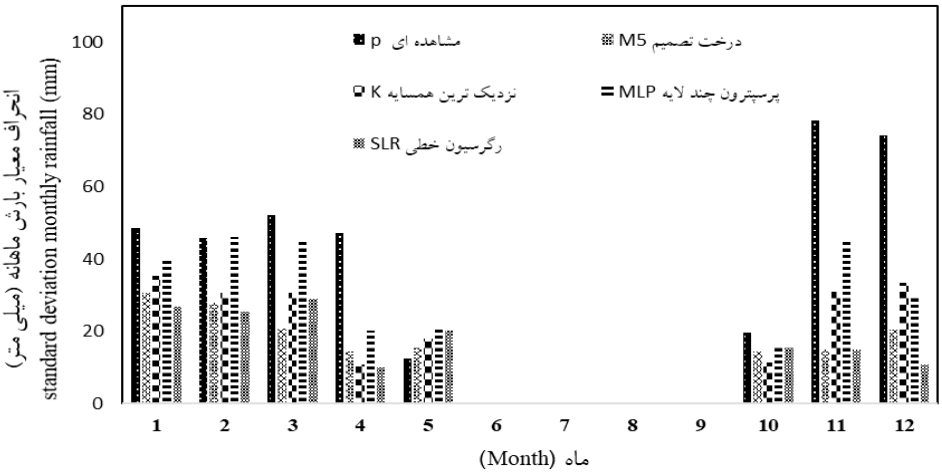
شکل ۲- مقادیر مشاهده‌شده و پیش‌بینی‌شده میانگین بارش ماهانه (بر حسب میلی‌متر) به روش **KNN** (بالا سمت راست)، روش **M5** (بالا سمت چپ) و روش **MLP** (پایین سمت راست)، روش **SLR** (پایین سمت چپ) در مقابل خط نیمساز.

Figure 2. Observed and predicted values of average monthly rainfall (in mm) by KNN (top right), M5 (top left) and MLP (bottom right), SLR (bottom left) in front of the bisecting line.



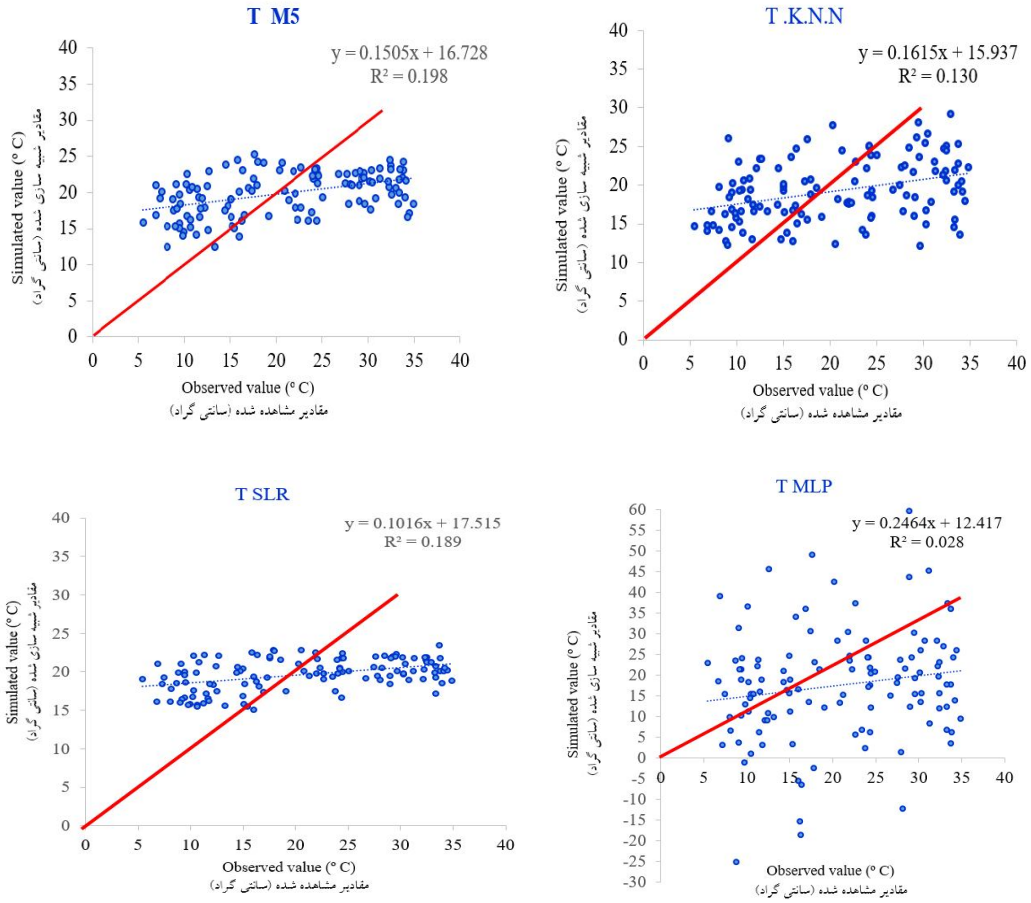
شکل ۳- میانگین بارش ماهانه مشاهده‌شده و پیش‌بینی‌شده توسط مدل‌ها در دوره آماری ۲۰۰۰-۱۹۹۱.

Figure 3. Observed and predicted average monthly rainfall by the models in the period of 1991-2000.



شکل ٤- انحراف معیار بارش ماهانه مشاهده‌شده و پیش‌بینی‌شده توسط مدل‌ها در دوره آماری ۲۰۰۰-۱۹۹۱.

Figure 4. Standard deviation of monthly precipitation observed and predicted by models in the period of 1991-2000.



شکل ٥- مقادیر مشاهده‌شده و پیش‌بینی‌شده میانگین دمای ماهانه (درجه سانتی‌گراد) به روش KNN (بالا سمت راست)، روش M5 (بالا سمت چپ) و روش MLP (پایین سمت راست)، روش SLR (پایین سمت چپ) در مقابل خط نیمساز.

Figure 5. Observed and predicted values of the average monthly temperature (C°) by KNN (top right), M5 (top left) and MLP (bottom right), SLR (Bottom left) in front of the bisecting bisecting line.

نتایج تحلیل‌های آماری با توجه به جدول ۲ نیز نشان داد که مدل درخت تصمیم در مرحله آزمون با توجه به معیارهای ریشه میانگین مربعات خطا و میانگین خطای اریب و ضریب همبستگی نسبت به مدل‌های دیگر، برآورد بهتری برای بارش و دمای ماهانه داشته است. اگرچه نتایج ضریب تعیین این مدل در مرحله آزمون برای برآورد دمای ماهانه، ضعیف‌تر از بارش ماهانه می‌باشد.

جدول ۲- نتایج تحلیل آماری مدل‌های پیش‌بینی‌کننده بارش و دمای ماهانه.

Table 2. Statistical analysis of prediction models for monthly temperature and rainfall.

معیار Criterion	نوع متغیر Variable Type	نزدیک‌ترین همسایه K-Nearest Neighbor		مدل درختی decision tree		پرسپترون چندلایه Multilayer Perceptron Network		رگرسیون خطی ساده Simple Linear Regression	
		آموزش Train	آزمون Test	آموزش Train	آزمون Test	آموزش Train	آزمون Test	آموزش Train	آزمون Test
ریشه میانگین مربعات خطا Root Mean Square Error	بارش Rainfall	36.41	53.5	36.90	46.87	28.53	66.71	36.60	52.28
	دما Temperature	8.62	10.44	9.02	8.97	6.09	17.93	10.26	10.27
میانگین خطای اریب Mean Bias Error	بارش Rainfall	20.30	35.46	27.57	32.71	21.07	43.62	1.39	42.33
	دما Temperature	8.37	10.16	8.58	8.67	5.95	13.19	11.26	13.50
ضریب همبستگی Correlation coefficient	بارش Rainfall	0.85	0.60	0.78	0.70	0.80	0.50	0.62	0.58
	دما Temperature	0.50	0.40	0.40	0.46	0.75	0.17	0.25	0.43

نتیجه‌گیری کلی

دما و بارش به‌عنوان دو متغیر مهم هواشناسی، به‌خصوص در مناطق خشک و نیمه‌خشک مطرح هستند. در نتیجه تعیین میزان این متغیرها، تغییرات آن‌ها و پیش‌بینی این پدیده‌ها به‌منظور برنامه‌ریزی دقیق‌تر در مدیریت بخش‌های کشاورزی، اقتصادی و اجتماعی ضروری می‌باشد. همچنین، بسیاری از سامانه‌های محیط زیست مانند منابع آب به‌علت وقوع پدیده تغییر اقلیم تحت‌تأثیر قرار گرفته‌اند؛ به‌طوری‌که بهره‌برداری از مخازن آب، تولید محصولات کشاورزی، فرسایش، تولید رواناب و بسیاری دیگر از فرایندهای هیدرولوژیکی دچار دگرگونی شده‌اند (۱۶). در این پژوهش، کارایی چهار مدل KNN، M5، SLR و MLP در مدل‌سازی بارش و دمای ماهانه ایستگاه هواشناسی مرغک با داده‌های خروجی مدل NCEP ارزیابی شد که بیانگر ضعف این مدل‌ها در ریزمقیاس‌نمایی بارش و دمای ماهانه بود. بنابراین با وجود برتری نسبی مدل درخت تصمیم M5 نسبت به سایر مدل‌ها، استفاده از مدل‌های داده‌کاوی مذکور برای پیش‌بینی متغیرهای بارش و دما در ایستگاه

مرغک توصیه نمی‌شود. هر چند به‌طورکلی، مدل درخت تصمیم به لحاظ سادگی و ایجاد قوانین پیش‌بینی بر مدل‌های جعبه سیاه برتری دارد. از طرفی، مدل مذکور قادر است بدون دخالت کاربر، ورودی‌های مهم‌تر را برای ایجاد قوانین پیش‌بینی استفاده و ورودی‌های ضعیف‌تر را حذف نماید.

منابع

1.Aksornsingchai, P., and Srinilta, CH. 2011. Statistical downscaling for rainfall and temperature prediction in Thailand. Proceeding of the international multi conference of engineers and computer scientists, 6p.

2.Chen, H., Yu Xu, C., and Guo, S. 2012. Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. J. Hydrol. 434-435: 36-45.

3.Dawson, C.W., and Wilby, R. 1988. An artificial neural network approach torainfall–runoff modeling. J. Hydrol. 43: 47-66.

4.Deepashree, R., and Mujumdar, P. 2011. A comparison of three methods for downscaling daily precipitation in the Punjab region. Hydrological Processes. 25: 23. 3575-3589.

5.Dibike, B.Y., and Coulibaly, P. 2006. Temporal neural networks for downscaling climate variability and extremes. J. Neur. Net. 19: 135-144.

6.Ghamghami, M. 2010. Evaluation and comparison of parametric models and nonparametric analysis of meteorological data. M.Sc. Thesis, Tehran University, 128p. (In Persian)

7.Ghorbani, Kh. 2015. Evaluation data mining models in Downscaling of precipitation based on NCEP general circulation model output (Case study: Kermanshah synoptic station). Iran Water Res. J. (IWRJ). 15: 15. 177-186. (In Persian)

8.Hamlet, A.F., and Lettenmaier, D.P. 2007. Effects of 20th century warming and climate variability on flood risk in the western U.S. Water Resources Research. 43: W06427, doi.1029/2006WR005099.

9.IPCC. 2007. Summary for Policymakers. P 1-18, In: S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (Eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge.

10.Khan, M.S., Coulibaly, P., and Dibike, Y. 2006. Uncertainty analysis of statistical downscaling methods. J. Hydrol. 319: 4. 357-382.

11.Kutner, M., Nachtcheim, Ch., and Neter, J. 2005. Applied Linear Statistical Models. McGraw-Hill Irvin Press, 1396p.

12.Meshkavati, A.M., kordjazi, M., and Babaeian, I. 2011. Evaluation of Lars models to simulate meteorological data Golestan Province in the period (1993-2007). J. Appl. Res. Geograph. Sci. 19: 81-96. (In Persian)

13.Mitchell, T.D. 2003. Pattern Scaling: An Examination of Accuracy of the Technique for Describing Future Climates. Climatic Change. 60: 217-242.

14.Quinlan, J.R. 1992. Learning with continuous classes. Proceedings of Fifth Australian joint conference on artificial intelligence, Singapore, Pp: 343-348.

15.Semenov, M.A., and Barrow, E.M. 2002. LARS-WG a stochastic weather generator for use in climate impact studies. User's manual, Version 3.0.

16.Seyyed Kaboli, H., Akhondali, A.M., Masah Bavani, A.R., and Radmanesh, F. 2012. A Downscaling Model Based on K-nearest neighbor (K-NN) Non-parametric Method. J. Water Soil. 26: 4. 779-808. (In Persian)

17.Tripathi, S., Srinivas, V., and Nanjundiah, R.S. 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. J. Hydrol. Pp: 621-640.

18.Two Crows Corporation. 1999. Introduction to datamining and knowledge discovery, third edition Available at: www.twocrows.com. 36p.
19.Zahoor, J., Abrar, M., Bashir, Sh., and Mirza, A. 2009. Seasonal to inter-annual climate prediction using data mining KNN technique. Wireless Networks, Information Processing and Systems, Communications in Computer and Information Science. 20: 40-51.
20.Witten, I.H., and Frank, E. 2005. Data mining practical machine learning tools and techniques with Java implementations. Morgan Kaufmann San Francisco, 664p.

Gorgan University of Agricultural Sciences and Natural Resources

J. of Water and Soil Conservation, Vol. 24(5), 2018
http://jwsc.gau.ac.ir

Comparative comparison of data mining models in downscaling rainfall and temperature (Case study: Bazoft-e-Samsami Watershed)

N. Dehghani[1], *H. Ghasemieh[2], S.J. Sadatinejad[3], Kh. Ghorbani[4] and A.A. Besalatpour[5]

[1]Ph.D. Student, Dept. of Range and Watershed Management, University of Kashan,
[2]Assistant Prof., Dept. of Range and Watershed Management, University of Kashan,
[3]Associate Prof., Dept. of Renewable Energies and Environment, University of Tehran,
[4]Associate Prof., Dept. of Water Engineering, Gorgan University of Agricultural Sciences and Natural Resources,
[5]Assistant Prof., Dept. of Soil Science, Vali-e-Asr University of Rafsanjan
Received: 11/15/2016; Accepted: 12/30/2017

Abstract

Background and Objectives: Temperature and rainfall are two important meteorological variables, especially in arid and semi-arid areas. As a result, determining the value of these variables, their changes and prediction of these phenomena are necessary for more precise planning in the management of agricultural, economic and social sectors. Nowadays, incompatibility of temporal and spatial scales required in investigated models on the effect of climate change with GCM outputs and the need to assess the change trend in meteorological threshold variables at the regional scale has led to develop various downscaling methods. So, the aim of this study is the comparative comparison of data mining models in downscaling of rainfall and temperature based on data of NCEP general circulation model.

Materials and Methods: The study area in this research is bazoft-e-Samsami watershed. This basin is one of the northern Karun sub-basins located in the northwest of Chaharmahal and Bakhtiari province. Marghak rain gauge and hydrometric stations are located at its outlet. In this study, the performance and efficiency of four methods including decision tree (M5), Nearest Neighbor (KNN), Multilayer Perceptron (MLP) and Simple linear regression (SLR) were evaluated for modeling monthly rainfall and temperature of Marghak station during the training period of 1971-1990 and the 1991-2000 test period using NCEP output parameters.

Results: Monthly rainfall modeling results using mentioned models showed that the output of all models except the KNN model provides negative values for rainfall. The rainfall prediction by M5 model in January, March, April and December is lower than the observed values (P). This situation is also somewhat seen in other models. Also, given that the minimum rainfall is zero, it can be concluded from the low predicted values rather than observed values that the maximum limit of rainfall with these models is not well predicted. The prediction of rainfall by all models in all months except May has a lower standard deviation than the observed values (P). The predicted results of monthly temperature also showed that only MLP output provides negative values for the temperature, which can be due to the extrapolation and generalizationin in MLP method. Also, The standard deviation obtained from all models in January, February, March, April, July, August, October, November and December is more than standard deviation of observed temperature. The results of statistical analyzes also showed that M5 than the other models in the test stage according to RMSE, MBE and R^2 have better estimates for rainfall and monthly temperature. Although the results of determination coefficient (R^2) in the test stage for monthly temperature estimation are weaker than monthly rainfall.

Conclusion: The results of the efficiency of four models of KNN, M5, SLR and MLP in monthly rainfall and temperature modeling in Marghak meteorological station with NCEP output data showed that these models were weak in downscaling the monthly rainfall and temperature. Therefore, despite the relative superiority of M5 model compared to other models, the use of these data mining models is not recommended to predict rainfall and temperature variables in Margak station.

Keywords: Downscaling, Decision tree (M5), Nearest neighbor (KNN), Multilayer perceptron (MLP), Simple linear regression

* Corresponding Author; Email: h.ghasemieh@kashanu.ac.ir