فیزیک زمین و فضا، دوره ۴۷، شماره ۲، تابستان ۱۴۰۰، صفحهٔ ۳۱۵- ۳۳۲ (مقاله پژوهشی) DOI: 10.22059/jesphys.2021.314958.1007269

بررسی کارایی روش‌های بازسازی نواقص آماری در رابطه با پارامتر بارش در مناطق خشک ایران

محمدرضا کوثری[1*]، میترا السادات اسمعیل‌زاده حسینی[2] و مرتضی میری[1]

۱. استادیار، پژوهشکده حفاظت خاک و آبخیزداری، سازمان تحقیقات، آموزش و ترویج کشاورزی، تهران، ایران
۲. دانشجوی دکتری، گروه مدیریت مناطق خشک و بیابانی، دانشکده منابع طبیعی و کویرشناسی، دانشگاه یزد، یزد، ایران

(دریافت: ۹۹/۱۱/۱۷، پذیرش نهایی: ۱۴۰۰/۳/۴)

چکیده

نواقص آماری عاملی رایج در داده‌های اقلیمی هستند و برای تخمین آنها تا به‌حال روش‌های متنوعی توسعه یافته‌اند. در این میان، روش‌های نسبت‌نرمال، رگرسیون خطی، رگرسیون چندمتغیره و عکس مجذور فاصله یا IDW از کاربرد گسترده‌ای در مطالعات منابع طبیعی کشور ما برخوردار هستند. در پژوهش حاضر، قابلیت هر یک از روش‌های مذکور در بازسازی نواقص آماری بارش روزانه، ماهانه و سالانه مناطق خشک کشور متناسب با میزان نسبت نقص داده از ۵ درصد تا ۵۰ درصد داده‌ها موردارزیابی قرارگرفت. نتایج نشان داد که هر روش متناسب با شرایط میزان داده‌های گم‌شده از عملکرد متفاوتی برخوردار است. روش رگرسیون چند متغیره هنگامی که نقص داده‌ها زیاد نباشد از دقت بیشتری در بازسازی داده‌های روزانه برخوردار است ولی در کل به میزان نسبت داده‌های گم‌شده حساس می‌باشد. روش نسبت‌نرمال در بازسازی نواقص بارش روزانه مناسب نیست ولی نسبت به میزان نقص داده‌ها از سایر روش‌ها پایدارتر است. در سری‌های زمانی ماهانه عملکرد IDW و سپس نسبت نرمال مناسب است. در سری‌های سالانه به‌ترتیب، روش همبستگی خطی، نسبت‌نرمال و IDW عمکرد بهتری دارند. در کل هر روش متناسب با شرایط بایستی مورد استفاده قرارگیرد و پیشنهاد می‌شود برای بازسازی نواقص آماری، یک بسته نرم‌افزاری برای کل کشور ارائه شود.

واژه‌های کلیدی: نقص داده، نسبت نرمال، همبستگی خطی، بازسازی، بارش.

۱. مقدمه

بارش بخشی بسیار مهم از چرخه هیدرولوژی است و مطالعات هواشناسی و هیدرولوژی حوزه‌های آبخیز در ابتدا بستگی به کمیت و کیفیت داده‌های ثبت‌شده بارش و پراکندگی آن در حوزه مورد نظر و مناطق همجوار دارد. برای برنامه‌ریزی و طراحی این پروژه‌ها، مجموعه‌های کامل و قابل‌اطمینان از داده‌های اقلیمی و هیدرولوژیکی، مورد نیاز است. آنچه مشخص است تقریباً در تمامی موارد، داده‌های گم‌شده یا نواقص آماری وجود دارند. داده‌های ناقص می‌توانند به‌شدت بر کیفیت و کاربرد مدل‌ها تأثیرگذار باشند (کامواگا و همکاران، ۲۰۱۸). چنین نقص‌هایی در سری‌های زمانی اقلیمی به‌خاطر مشکلات تجهیزات و ادوات پایش پارامترهای هواشناسی، تأثیرات پدیده‌های طبیعی مثلاً زمین لرزه، طوفان، لغزش زمین، نقص در هنگام انتقال داده، ذخیره‌سازی و فرایندهای بازیابی مربوط می‌شود. همچنین عواملی مانند تعطیلی موقت ایستگاه‌ها، عدم‌حضور فرد داده‌بردار در زمان مورد نظر و یا کوتاهی دوره برداشت داده، همگی باعث ایجاد کوتاهی سری زمانی یا نواقص آماری می‌شوند(علیزاده، ۱۳۹۲؛ کانچالا نستار و همکاران، ۲۰۱۹؛ جودی و همکاران، ۱۳۹۵؛ مهدوی، ۱۳۸۴). در رابطه با بازسازی نواقص آماری، دو حالت کلی وجود دارد. اولین حالت نقص دوره آماری است به‌طوری‌که در میان داده‌های برداشت شده ایستگاه، موارد داده‌های گم‌شده نیز به‌چشم می‌خورد. در حالت بعدی، سری زمانی موردنظر، از طول دوره آماری کافی برخوردار نیست. لذا در این حالت نیاز به تطویل طول سری‌های زمانی می‌باشد. به‌هرحال، بسیاری از مدل‌های محیطی نیاز به داده‌های کامل و بدون نقص دارند و لذا بازسازی نواقص آماری نقش بسیار مهمی در کارکرد این مدل‌ها دارند (اوبی و همکاران، ۲۰۰۰؛ بارلو و همکاران، ۲۰۱۶؛ هو و هوانگ، ۲۰۲۰).

به‌منظور بازسازی نواقص آماری، روش‌های بسیار متنوعی

*نگارنده رابط: mohammad_kousari@yahoo.com

توسعه یافته است. انتخاب نوع روش به عواملی نظیر طول سری زمانی داده‌های ناقص، در دسترس بودن داده‌های آب‌وهواشناسی از ایستگاه‌های همجوار، فصل مقادیر گمشده، منطقه آب‌وهوایی موردنظر، دانش و تخصص شخص مسئول تصحیح داده‌ها، طول ضبط داده‌های موجود، اهمیت کیفیت داده برای مدل موردنظر بستگی دارد. بدین‌منظور روش‌های مختلفی نظیر رگرسیون خطی ساده و چند متغیره (کامواگا و همکاران، ۲۰۱۸؛ ستاری و همکاران، ۲۰۱۶)، شبکه‌های عصبی مصنوعی (کانچالا نستار و همکاران، ۲۰۱۹؛ کیم و پاچسکی، ۲۰۱۰)، روش‌های فازی (اوبی و همکاران، ۲۰۰۰)، درخت تصمیم (کیم و پاچسکی، ۲۰۱۰)، تحلیل سری‌های زمانی و تکنیک‌های درون‌یابی (فوهن و همکاران، ۲۰۱۸؛ تیگاورپو، ۲۰۲۰)، روش نسبت نرمال (حسن‌پور کاشانی و دینپشوه، ۲۰۱۲)، روش‌های مبتنی بر تحلیل فضایی داده‌ها مانند روش عکس مجذور فاصله یا IDW (Inverse Distance Weighting) و کریجینگ (اشتیلی‌نوا و همکاران، ۲۰۱۷) به‌کار گرفته می‌شوند. البته در کشور ما و در مطالعات منابع طبیعی و به‌جهت اینکه در کتاب‌های پایه آموزش هیدرولوژی و اقلیم‌شناسی (علیزاده، ۱۳۹۲؛ مهدوی، ۱۳۸۴) بیشتر به رویکردهای ساده و کاربردی پرداخته شده است، روش‌هایی مانند رگرسیون خطی ساده و چند متغیره، نسبت نرمال و IDW بیشترین کاربرد را دارا می‌باشند. در واقع از این روش‌ها در بسیاری از مطالعات پایه اقلیم‌شناسی خصوصاً در رابطه با مطالعات حوزه‌های آبخیز استفاده می‌شود.

تا به‌حال مطالعات زیادی در رابطه با بررسی دقت و عملکرد روش‌های بازسازی نواقص آماری انجام شده است. در ایستگاه سینوپتیک سراب، روش‌های متعدد بازسازی داده‌ها، با استفاده از هفت روش کلاسیک آماری و مدل درختی M5 موردبررسی قرارگرفت. نتایج این تحقیق نشان داد که از بین روش‌های کلاسیک آماری بررسی شده به‌ترتیب روش‌های انتصاب چندگانه، نسبت نرمال و رگرسیون خطی چندمتغیره دارای عملکرد نسبتاً دقیق‌تر با خطاهای کمتری می‌باشند. همچنین مدل درختی M5 دقت بیشتری نسبت به سایر روش‌ها داشته است (جودی و همکاران، ۱۳۹۵). در پژوهشی دیگر روش‌های کلاسیک و همچنین مبتنی بر هوش مصنوعی برای بازسازی نواقص آماری داده‌ها موردارزیابی قرار گرفت (ریس، ۲۰۰۸). نتایج این تحقیق نشان می‌دهد که هرچند رویکرد مربوط به هوش مصنوعی با پیچیدگی و محاسبات بیشتری برای پیدا کردن ساختارهای بهینه همراه هستند، اما دقت بیشتری نسبت به روش‌های کلاسیک دارند. نتایج بازسازی داده بارش ماهانه در شش ایستگاه با استفاده از روش‌های میانگین حسابی، رگرسیون چند متغیره و NLTPLS (Non-Linear Iterative Partial Least Squares) در استان هرمزگان (ستاری و همکاران، ۲۰۱۶) نشان می‌دهد که روش رگرسیون چندمتغیره از دقت بیشتری برخودار است. پژوهش دیگری در استان اصفهان به‌منظور بازسازی داده‌های بارش در اقلیم‌های متفاوت انجام گرفت. روش‌های مورد بررسی در تحقیق مذکور شامل روش‌های ایستگاه معرف، نسبت نرمال، محور مختصاتی، رگرسیون خطی، رگرسیون چندمتغیره و سری زمانی می‌باشد. یافته‌ها حاکی از آن است که در اقلیم خشک استان اصفهان روش نسبت نرمال، در اقلیم نیمه‌خشک روش رگرسیون چندمتغیره و در اقلیم مدیترانه‌ای روش نسبت نرمال بهترین بازسازی را برای داده‌های بارش سالانه ارائه کرده است (ساداتی‌نژاد و همکاران، ۱۳۷۶). در مطالعه دیگری (تیگاورپو، ۲۰۲۰) روش اندازه‌گیری احتمالاتی خطا بر مبنای فاصله برای بازسازی داده‌های بارش مورد بررسی قرار گرفت. نتایج نشان دهنده عملکرد بهتر این رویکرد نسبت به استفاده از فاصله اقلیدسی برای بازسازی داده‌ها می‌باشد. بررسی کارایی روش‌های مبتنی بر یادگیری ماشین در بازسازی داده‌های ناقص نشان داد که شیوه بهینه‌سازی خطی و شبکه‌های عصبی مصنوعی در بازسازی نواقص داده‌های بارش مناسب هستند (تیگاورپو و همکاران، ۲۰۱۸). نتایج پژوهشی دیگر برای بازسازی نواقص آماری نشان می‌دهد

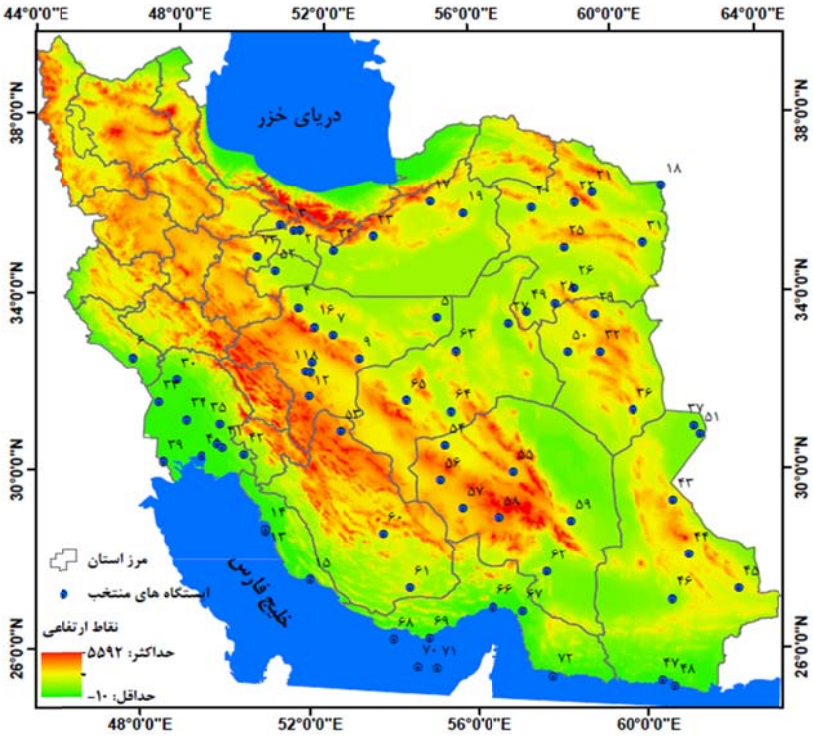
که از بین رویکردهای جایگزینی ساده، همبستگی خطی و نمایی و رگرسیون رتبه‌ای، روش جایگزینی ساده با میانگین داده‌ها و همبستگی نمایی عملکرد بهتری دارد (لوپرستی و همکاران، ۲۰۰۸).

بر اساس مطالعات انجام شده تاکنون روش‌های مختلف و متنوعی برای بازسازی نواقص آماری داده‌های هواشناسی و از جمله بارش توسعه یافته‌اند و عملکردهای متفاوتی از هر روش ارائه شده است. هر روش مزیت‌ها و معایب خود را دارد ولی به‌هر حال، بایستی کارایی رویکردهای مختلف بازسازی داده خصوصاً موارد پرکاربرد آن مشخص شود. محققان و کارشناسان همواره با سری‌های زمانی مختلف بارش مانند سری‌های روزانه، ماهانه و سالانه سر و کار دارند، اما همواره این سوال باقی است که کدام روش، در چه شرایطی و برای چه سری زمانی مناسب است. هدف اصلی تحقیق حاضر، پاسخ به این سوال است که در مناطق خشک ایران که پراکنش مکانی و زمانی بارش بسیار زیاد است، کدام یک از چهار روش موردمطالعه شامل روش نسبت نرمال، رگرسیون خطی، رگرسیون چندمتغیره IDW و برای کدام یک از سری‌های زمانی مناسب‌تر است. خصوصاً اینکه میزان نقص در داده‌های اولیه به چه میزان بر عملکرد روش تأثیرگذار است.

۲. روش پژوهش

۲-۱. منطقه مورد مطالعه

منطقه موردمطالعه، نواحی خشک ایران است. ایران دارای گستردگی مساحت و همچنین تغییرات توپوگرافی قابل‌توجهی است. لذا اقلیم‌های متنوعی در ایران وجود دارد. بر اساس تقسیم‌بندی انجام‌شده به روش کوپن_گایگر از ۳۱ گروه اقلیمی موجود در این روش، هشت گروه اقلیمی آن در ایران وجود دارد که منطقه موردمطالعه در گروه اقلیم بیابانی خشک و گرم (BWh) قرا می‌گیرد (رضیئی، ۱۳۹۶)،که بخش قابل‌توجهی از کشور را تشکیل می‌دهد. به‌خصوص مناطق بعد از ارتفاعات البرز و زاگرس در این محدوده قرار می‌گیرند. این ارتفاعات نقش بسیارمهمی در تعیین نوع اقلیم ایران و حتی قسمت‌های قابل‌توجهی از خاورمیانه دارند (بارلو و همکاران، ۲۰۱۶).



شکل ۱. پراکندگی مکانی ایستگاه‌های مورداستفاده در تحقیق حاضر.

۲-۲. داده‌های مورداستفاده

از داده‌های روزانه ایستگاه‌های هواشناسی سینوپتیک کل کشور برای تحقیق حاضر استفاده شد. با استفاده از آزمون همگنی ران تست، همگنی داده‌ها بررسی شد. همچنین با استفاده از روش‌های ترسیم نمودار و به‌خصوص نمودارهای جعبه‌ای، داده‌های پرت نیز مشخص شده و به داده‌های ناقص تبدیل شدند. میانگین بارش و دمای سالانه ۴۰۰ ایستگاه تعیین و سپس بر اساس این داده‌ها، ضریب دومارتن (دومارتن، ۱۹۲۵) آنها استخراج شد. در مرتبه بعدی ایستگاه‌های دارای ضریب دومارتن کمتر از ۱۰ به‌عنوان مناطق دارای اقلیم خشک انتخاب شدند. از میان آنها، تعداد ۷۳ ایستگاهی که از سال ۱۳۶۵ تا ۱۳۹۶ دارای آمار کافی بودند، متمایز شدند. برای ارزیابی هر یک از روش‌های بازسازی داده، بخشی از داده‌های واقعی به‌صورت تعمدی، از داده‌های اصلی کنار گذاشته و سپس مورد بازسازی قرار گرفت. لذا سعی بر آن شد که سری‌های زمانی ایستگاه‌های مورد بررسی قرار گیرد که خود دارای حداقل نقص آماری هستند.

نقشه پراکندگی ایستگاه‌های مذکور در ایران در شکل ۱ نشان داده شده است. لازم به‌ذکر است که مشخصات ایستگاه‌ها برحسب شماره ردیف در جدول ۱ ارئه شده است. مشخصات ایستگاه‌های مورد بررسی در تحقیق حاضر در جدول ذیل (جدول ۲) ارائه شده است. عدد ارائه‌شده برای هر ایستگاه در نقشه با مقدار مشخص شده در ستون ردیف این جدول همخوانی دارد.

جدول ۱. تقسیم‌بندی ضریب دومارتن.

نام اقلیم	خشک	نیمه‌خشک	مدیترانه‌ای	نیمه مرطوب	مرطوب	بسیار مرطوب
ضریب خشکی دومارتن	کمتر از ۱۰	۱۰ تا ۱۹/۹	۲۰ تا ۲۳/۹	۲۴ تا ۲۷/۹	۲۸ تا ۳۴/۹	بیشتر از ۳۵

جدول ۲. مشخصات ایستگاه‌های مورداستفاده.

ردیف	نام ایستگاه	متوسط بارش (mm)	متوسط دما (C)	شاخص خشکی	ردیف	نام ایستگاه	متوسط بارش (mm)	متوسط دما (C)	شاخص خشکی
۱	کرج	۲۴۶/۸۸	۱۵/۳۳	۹/۷۴	۳۸	امیدیه (پایگاهی)	۲۲۷/۲۴	۲۵/۷۵	۶/۳۶
۲	دوشان تپه	۲۵۰/۱۸	۱۸/۱۲	۸/۹۰	۳۹	آبادان	۱۶۲/۱۴	۲۵/۸۳	۴/۵۲
۳	تهران مهرآباد	۲۳۰/۳۹	۱۷/۷۲	۸/۳۱	۴۰	بندر ماهشهر	۱۹۲/۵۲	۲۵/۹۲	۵/۳۶
۴	کاشان	۱۴۵/۹۵	۱۹/۲۴	۴/۹۹	۴۱	امیدیه (آغاجاری)	۲۵۴/۱۷	۲۶/۰۹	۷/۰۴
۵	خور و بیابانک	۷۶/۵۹	۲۰/۷۰	۲/۴۹	۴۲	بهبهان	۳۲۳/۲۷	۲۴/۹۸	۹/۲۴
۶	دهلران	۲۸۲/۸۵	۲۶/۲۳	۷/۸۱	۴۳	زاهدان	۸۰/۱۹	۱۸/۶۳	۲/۸۰
۷	اردستان	۱۲۸/۹۳	۱۹/۱۷	۴/۴۲	۴۴	خاش	۱۴۲/۵۴	۲۰/۲۶	۴/۷۱
۸	اصفهان (ازن سنجی)	۱۲۲/۳۲	۱۶/۵۷	۴/۶۰	۴۵	سراوان	۱۰۶/۰۲	۲۲/۲۸	۳/۲۸
۹	نائین	۹۷/۷۱	۱۶/۸۵	۳/۶۴	۴۶	ایرانشهر	۱۰۷/۵۵	۲۶/۹۵	۲/۹۱
۱۰	اصفهان (فرودگاه)	۱۰۲/۱۹	۱۵/۱۵	۴/۰۶	۴۷	کنارک (فرودگاه)	۱۰۷/۵۸	۲۶/۷۰	۲/۹۳
۱۱	کبوتر آباد	۱۱۳/۸۲	۱۵/۶۲	۴/۴۴	۴۸	چابهار	۱۱۶/۴۷	۲۶/۳۴	۳/۲۰
۱۲	شهر ضا	۱۳۸/۹۹	۱۴/۷۵	۵/۶۲	۴۹	بشرویه	۸۵/۷۹	۱۹/۴۲	۲/۹۲
۱۳	بوشهر (ساحلی)	۲۵۳/۹۳	۲۵/۲۶	۷/۲۰	۵۰	خور بیرجند	۸۴/۱۹	۱۹/۹۴	۲/۸۱
۱۴	بوشهر (فرودگاه)	۲۳۷/۳۴	۲۴/۹۰	۶/۸۰	۵۱	زهک	۴۸/۴۹	۲۳/۰۵	۱/۴۷
۱۵	بندر دیر	۲۳۱/۵۴	۲۷/۴۴	۶/۱۸	۵۲	قم	۱۳۹/۱۳	۱۸/۴۱	۴/۹۰
۱۶	نطنز	۱۸۴/۸۹	۱۵/۸۵	۷/۱۵	۵۳	آباده	۱۳۴/۰۳	۱۴/۴۶	۵/۴۸

ردیف	نام ایستگاه	متوسط بارش (mm)	متوسط دما (C)	شاخص خشکی	ردیف	نام ایستگاه	متوسط بارش (mm)	متوسط دما (C)	شاخص خشکی
۱۷	شاهرود	۱۶۶/۶۱	۱۴/۹۲	۶/۶۹	۵۴	انار	۷۰/۱۹	۱۸/۴۶	۲/۴۷
۱۸	سرخس	۱۸۸/۱۵	۱۸/۲۸	۶/۶۵	۵۵	کرمان	۱۳۶/۳۲	۱۵/۹۳	۵/۲۶
۱۹	بیارجمند	۱۲۰/۱۱	۱۶/۳۰	۴/۵۷	۵۶	شهر بابک	۱۴۸/۰۱	۱۵/۴۲	۵/۸۲
۲۰	سبز وار	۱۶۹/۳۰	۱۷/۷۹	۷/۰۶	۵۷	سیرجان	۱۳۶/۹۷	۱۷/۴۴	۴/۹۹
۲۱	گلمکان	۲۱۴/۱۶	۱۳/۷۶	۹/۰۲	۵۸	بافت	۲۵۰/۹۲	۱۵/۲۴	۹/۹۴
۲۲	نیشابور	۲۳۴/۷۵	۱۴/۵۴	۹/۵۷	۵۹	بم	۵۸/۹۱	۲۳/۳۸	۱/۷۶
۲۳	سمنان	۱۳۷/۹۵	۱۸/۳۳	۴/۸۷	۶۰	فسا	۲۸۸/۰۶	۱۹/۴۸	۹/۷۷
۲۴	گرمسار	۱۱۲/۰۶	۱۹/۰۹	۳/۸۵	۶۱	لار	۲۰۴/۸۴	۲۳/۸۷	۶/۰۵
۲۵	کاشمر	۱۹۳/۲۹	۱۷/۹۹	۶/۹۱	۶۲	کهنوج	۱۸۲/۸۶	۲۶/۹۴	۴/۹۵
۲۶	گناباد	۱۳۱/۴۸	۱۷/۵۹	۴/۷۶	۶۳	رباط پشت بادام	۱۰۵/۶۴	۱۹/۲۲	۳/۶۱
۲۷	طبس	۷۹/۶۵	۲۲/۲۱	۲/۴۷	۶۴	بافق	۴۶/۸۶	۲۱/۶۵	۱/۴۸
۲۸	فردوس	۱۳۱/۷۰	۱۷/۵۷	۴/۷۸	۶۵	یزد	۶۷/۵۲	۱۹/۵۷	۲/۲۸
۲۹	قائن	۱۵۸/۹۵	۱۴/۷۸	۶/۴۱	۶۶	بندر عباس	۱۶۷/۹۸	۲۷/۰۳	۴/۵۴
۳۰	صفی آباد (دزفول)	۳۱۷/۵۹	۲۴/۲۶	۹/۲۷	۶۷	میناب	۱۹۸/۴۵	۲۷/۵۴	۵/۲۹
۳۱	تربت جام	۱۶۸/۹۸	۱۵/۸۹	۶/۵۳	۶۸	جزیره کیش	۱۵۳/۸۸	۲۷/۳۱	۴/۱۲
۳۲	بیرجند	۱۷۳/۵۷	۱۶/۵۶	۶/۵۳	۶۹	بندر لنگه	۱۳۲/۷۶	۲۶/۸۷	۳/۶۰
۳۳	بستان	۱۹۵/۱۰	۲۴/۶۶	۵/۶۳	۷۰	جزیره سیری	۱۰۴/۴۴	۲۷/۹۰	۲/۷۶
۳۴	اهواز	۲۱۶/۵۴	۲۵/۷۰	۶/۰۷	۷۱	جزیره ابوموسی	۱۱۸/۵۲	۲۷/۶۳	۳/۱۵
۳۵	رامهرمز	۲۸۵/۱۴	۲۶/۶۷	۷/۷۸	۷۲	جاسک	۱۲۵/۴۸	۲۷/۰۸	۳/۳۸
۳۶	نهبندان	۱۱۲/۷۵	۲۰/۱۳	۳/۷۴	۷۳	ساوه	۱۹۰/۴۵	۱۸/۵۴	۶/۶۷
۳۷	زابل	۳۱/۰۹	۲۲/۳۹	۱/۶۷	۷۴	میانگین کل	۱۵۹/۳۷	۲۰/۸۲	۵/۲۵

۳. اندازه‌گیری، مشاهده و محاسبه

روش‌های مرسوم بازسازی مورداستفاده در این پژوهش شامل؛ روش‌های نسبت نرمال، رگرسیون خطی، رگرسیون چندگانه و عکس مجذور فاصله (IDW) می‌باشند.

۳-۱. شیوه‌های بازسازی داده‌ها

در این روش، به ایستگاه‌های همجواری که از لحاظ میانگین پارامتر بارش به ایستگاه دارای نقص آماری نزدیک‌تر هستند، وزن بیشتری در بازسازی داده ناقص داده می‌شود. محاسبه این روش با رابطه ۱ صورت گرفت:

$$\frac{Px}{Nx} = \frac{1}{N}\sum_{i=1}^{N}\frac{Pi}{Ni} \qquad (۱)$$

که در آن Px مقدار نقص آماری بارشی است که بایستی تخمین زده شود. Nx میانگین بارش در ایستگاهی است که دارای نقص آماری است، یعنی ایستگاه هدف. N تعداد ایستگاه‌های همجوار یا ایستگاه شاهد که در بازسازی نقص آماری شرکت می‌کنند. Pi مقدار بارش در هر یک از ایستگاه‌های همجوار و همزمان با نقص آماری Px و Ni مقدار میانگین بارش در ایستگاه‌های همجوار است.

۳-۲. روش رگرسیون خطی

در این روش ابتدا همبستگی میان ایستگاه دارای نقص آمار با سایر ایستگاه‌ها از طریق رابطه ۲ تعیین می‌شود.

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} \qquad (۲)$$

در رابطه فوق، x و y به‌ترتیب سری‌های زمانی داده‌های بارش در ایستگاه شاهد و ایستگاه دارای نقص آماری هستند. از آنجایی‌که ماتریس همبستگی به‌صورت قرینه است، هرکدام از سری‌های زمانی در ایستگاه شاهد و هدف می‌توانند به‌عنوان x و y معرفی شوند. r میزان همبستگی بین ایستگاه را نشان می‌دهد که مقادیر آن بین ۱- تا ۱ متغیر است. n طول سری زمانی داده‌ها در ایستگاه شاهد یا هدف است. چنان‌چه ارتباط بین سری زمانی ایستگاه دارای آمار ناقص با برخی ایستگاه‌ها از لحاظ آماری معنادار بود، ایستگاهی به‌عنوان شاهد انتخاب می‌شود که قدر مطلق مقدار همبستگی آن بیشترین باشد. سپس از طریق برقراری یک رابطه خطی بین داده‌های ایستگاه شاهد و هدف، مقدار بارش در ایستگاه دارای نقص آماری از رابطه ۳ محاسبه شد.

$$y = a + bx \qquad (۳)$$

y در معادله فوق برابر با بارش در ایستگاه دارای نقص آماری و x مقدار بارش در ایستگاه شاهد و در زمانی است که نقص آماری در ایستگاه هدف وجود دارد. a و bنیز به‌ترتیب عرض از مبدأ و شیب خط در معادله درجه یک می‌باشند.

۳-۳. روش رگرسیون چند متغیره

این روش مشابه روش رگرسیون خطی است. البته معمولاً بیش از یک ایستگاه شاهد در بازسازی داده‌ها مشارکت می‌کند. در پژوهش حاضر، هر ایستگاهی که ضریب همبستگی آن از لحاظ آماری و در حد ۵ درصد با ایستگاه هدف معنادار بود، به‌عنوان ایستگاه شاهد در نظر گرفته شد.

۳-۴. روش عکس مجذور فاصله (IDW)

در این روش میزان نزدیکی بین ایستگاه‌ها عامل وزن و مشارکت ایستگاه‌های شاهد در بازسازی نقص داده ایستگاه هدف است. ابتدا وزن هر یک از ایستگاه‌ها با استفاده از (رابطه ۴) محاسبه شده است.

$$W = \frac{1}{X^2 + Y^2} \qquad (۴)$$

در رابطه ۴، X و Y به‌ترتیب فاصله طولی و عرضی بین هر ایستگاه شاهد و ایستگاه هدف است. بدیهی است هرچه فاصله کمتر باشد، میزان W یا وزن آن ایستگاه در بازسازی نواقص آماری ایستگاه هدف است بالاتر می‌رود. سپس بر اساس رابطه ۵، مقدار بارش در ایستگاه هدف P_X، و بر اساس وزن هر ایستگاه شاهد و بارش آن در تاریخی که نقص آماری وجود دارد (P_i)، تخمین زده می‌شود (آهنی و همکاران، ۲۰۱۲).

$$P_X = \frac{\sum W_i . P_i}{\sum W_i} \qquad (۵)$$

۳-۵. ارزیابی روش‌ها

از شاخص‌های ضریب همبستگی R (رابطه ۶)، یا معیار برآورد خطا RMSE (رابطه ۷) و ضریب Nash (رابطه ۸) برای ارزیابی دقت داده‌ها استفاده شد.

$$R = \frac{\sum PO - \frac{\sum P \sum O}{n}}{\sqrt{\left[\sum P^2 - \frac{(\sum P)^2}{n}\right]\left[\sum O^2 - \frac{(\sum O)^2}{n}\right]}} \qquad (۶)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \qquad (۷)$$

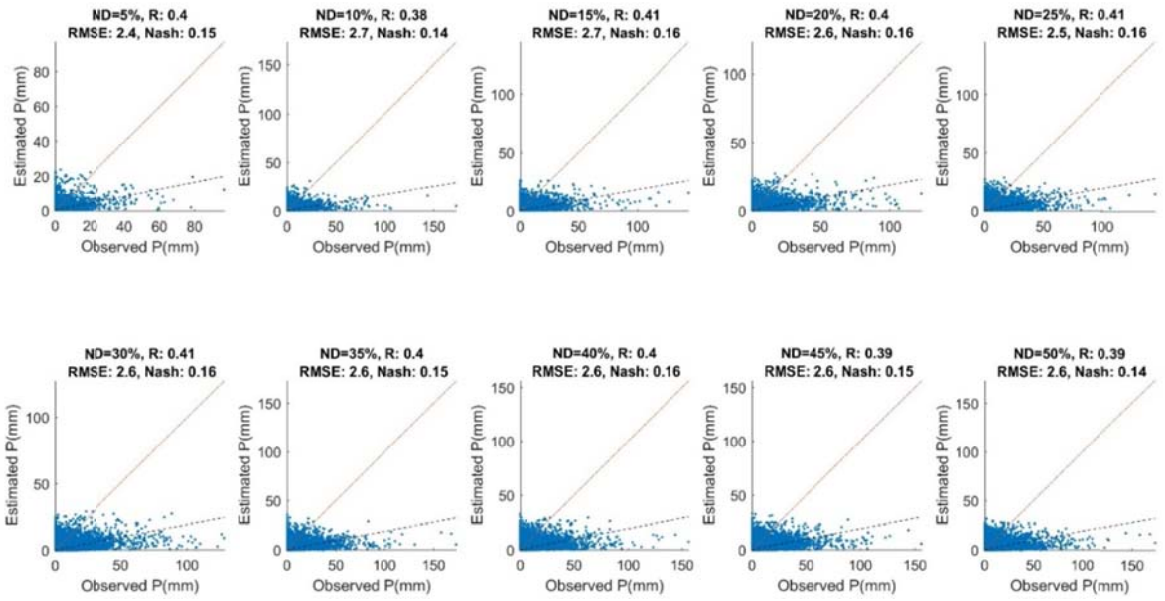
$$Nash = \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (O_i - O_{Avr})^2} \qquad (۸)$$

که در این روابط P و O به‌ترتیب مقادیر پیش‌بینی‌شده و مشاهده شده می‌باشند. O_{Avr}میانگین داده‌های مشاهداتی و n تعداد نمونه‌ها است (میری و همکاران، ۲۰۱۹). همچنین مقادیر واقعی و برآورد شده نیز بر روی یک نمودار ترسیم شد. در هر نمودار یک خط با زاویه ۴۵ درجه یا خط یک‌به‌یک قرار دارد. همچنین خط برازش بر نقاط نیز در هر نمودار نمایش داده شد. در صورتی‌که دقت برآوردها بیشتر باشد نقاط با همبستگی بیشتری اطراف خط مرکزی نمودار قرار می‌گیرند و لذا خط

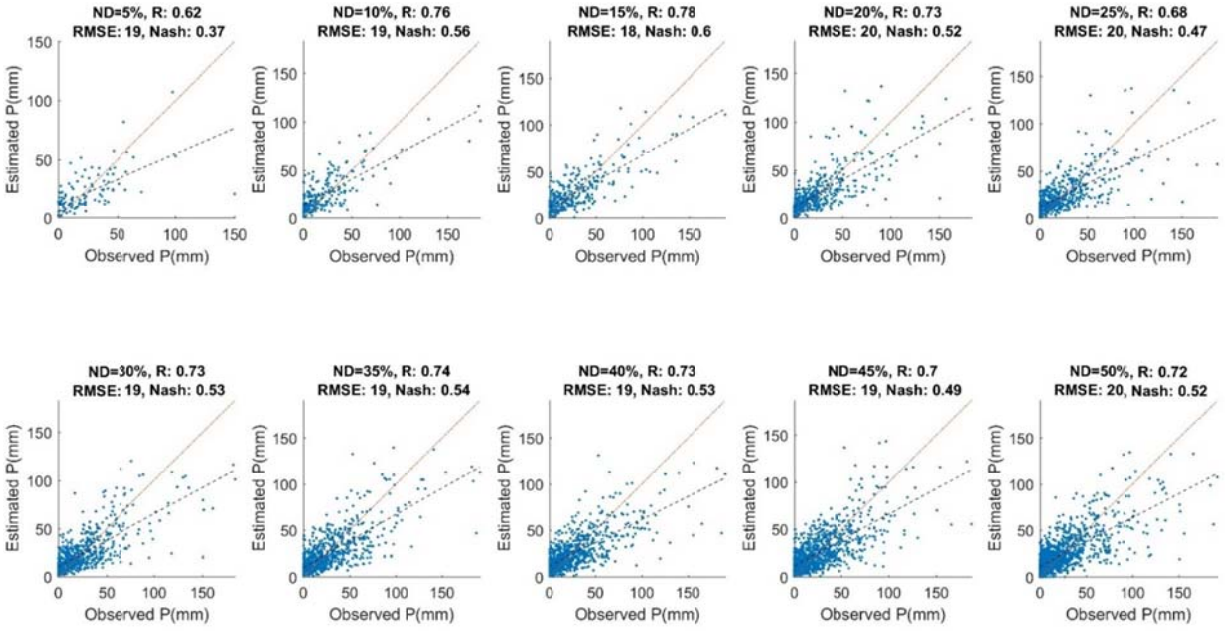
برازش داده شده بر داده‌های واقعی و برآورد شده نیز با خط یک‌به‌یک نمودار تطابق بیشتری دارد. برای ارزیابی روش‌های فوق‌الذکر، به‌صورت تعمدی، بخشی از داده‌های واقعی به داده ناقص یا No data تبدیل شدند. انتخاب آنها به‌صورت کاملاً تصادفی انجام شد. ایجاد داده‌های گم‌شده به‌گونه‌ای صورت گرفت که طی ۱۰ مرحله، از ۵ درصد تا ۵۰ درصد از داده‌های موجود به‌صورت داده ناقص درآمدند تا توان هر یک از روش‌ها در بازسازی داده‌ها متناسب با مقدار داده‌های گم‌شده نیز تعیین شود. در شرایط واقعی، گاهاً داده‌ها، دارای نسبت قابل‌توجهی از نقص هستند. لازم به‌ذکر است که تمامی محاسبات مربوط به بازسازی نواقص آماری به سه روش مذکور در تحقیق حاضر، با کمک برنامه‌نویسی در محیط نرم‌افزار MATLAB انجام شد.

۴. نتایج

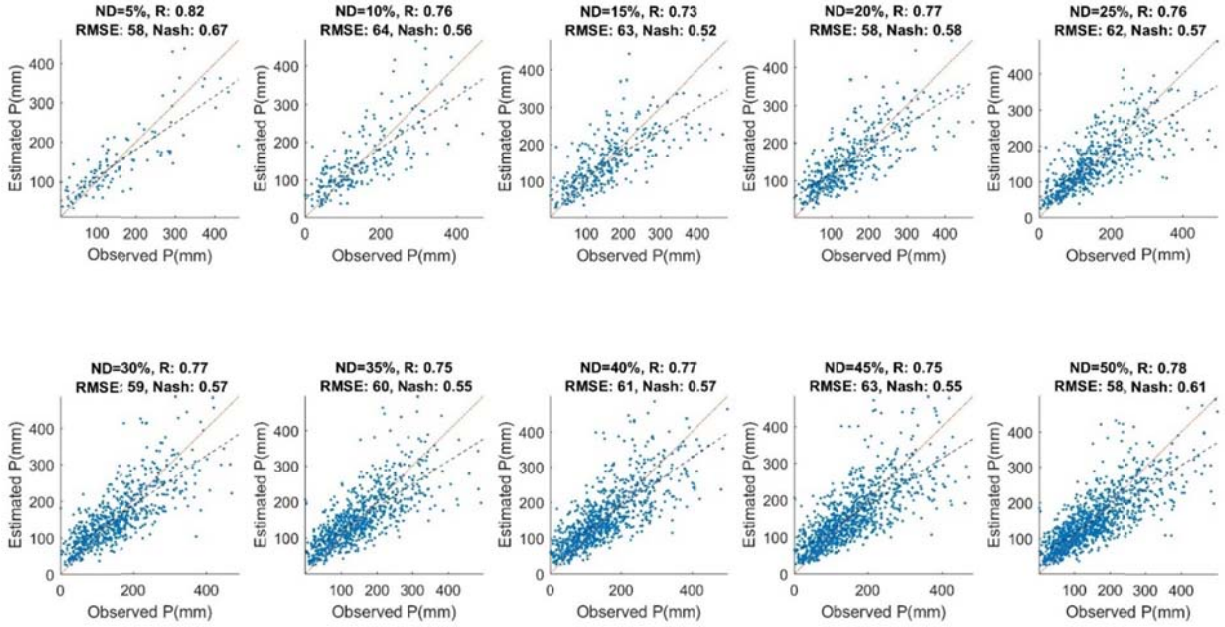
مقادیر بازسازی‌شده در مقابل مقادیر واقعی بررسی شد (شکل ۲ تا شکل ۱۳). هر شکل مربوط به یک روش موردبررسی و یک مقیاس زمانی یعنی روزانه، ماهانه و سالانه است و هر کدام شامل ۱۰ نمودار است که متناسب با درصد میزان نقص داده‌ها از ۵ تا ۵۰ درصد، پراکنش مقادیر بازسازی‌شده در مقابل مقادیر واقعی را نشان می‌دهد. در عنوان هر یک از نمودارها ND به‌معنای No Data است و مربوط به‌همین درصد ناقص شده از داده‌های واقعی است. همچنین در عنوان هر نمودار، مقدار R و RMSE و ضریب Nash نیز نمایش داده شده است. در محور xها مقادیر واقعی قرار دارند و محور yها مربوط به داده‌های بازسازی شده است. پراکنش مقادیر بازسازی شده با استفاده از روش نسبت نرمال در مقابل مقادیر واقعی بارش روزانه در شکل ۲ ارائه شده است. نتایج حاکی از آن است که داده‌های روزانه با استفاده از روش نسبت نرمال به‌خوبی برآورد نمی‌شوند و داده‌های گم‌شده را کمتر از حد واقعی برآورد می‌کند. پراکنش نقاط در رابطه با مقادیر بازسازی شده و واقعی بارش با استفاده از روش نرمال و در مقیاس زمانی ماهانه نمایش داده شده است (شکل ۳). در این نمودارها، پراکنش نقاط با نظم بهتری در اطراف خط با زاویه ۴۵ درجه دیده می‌شود. چنین وضعیتی در مقیاس سالانه نیز مشاهده می‌شود (شکل ۴).



شکل۲. مقادیر بازسازی‌شده داده‌های بارش روزانه با استفاده از روش نسبت نرمال در مقابل مقادیر واقعی بارش روزانه مقادیر داده‌های ناقص به‌صورت ساختگی از ۵ درصد تا ۵۰ درصددادده‌ها، ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



شکل۳. مقادیر بازسازی‌شده داده‌های بارش ماهانه با استفاده از روش نسبت نرمال در مقابل مقادیر واقعی بارش ماهانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.

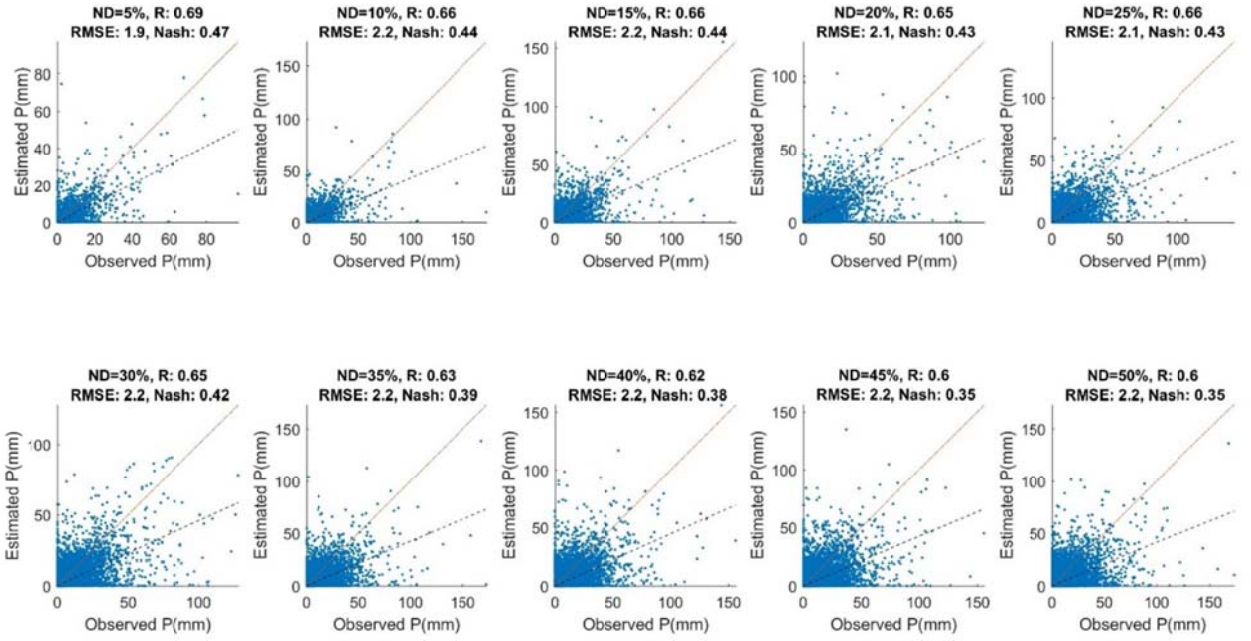


شکل۴. مقادیر بازسازی‌شده داده‌های بارش سالانه با استفاده از روش نسبت نرمال در مقابل مقادیر واقعی بارش سالانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.

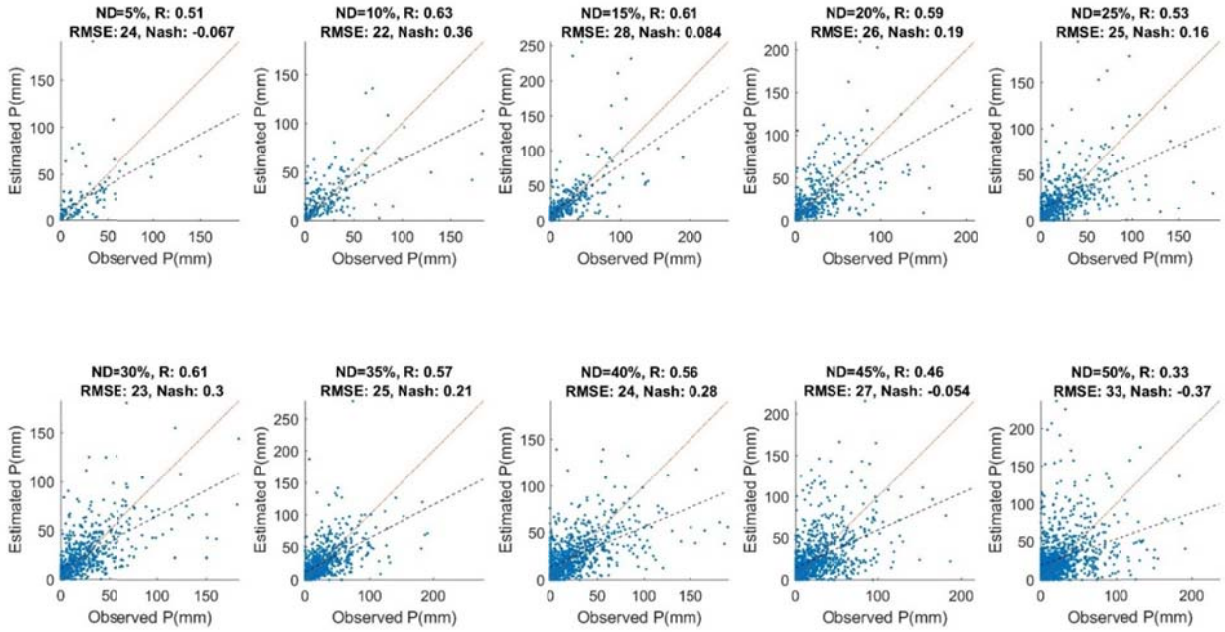
یافته‌های بعدی مربوط به پراکنش نقاط در رابطه با مقادیر بازسازی‌شده و واقعی بارش با استفاده از روش همبستگی خطی و در مقیاس زمانی روزانه (شکل ۵)، ماهانه (شکل ۶) و سالانه (شکل ۷) است. این شکل‌ها نشان می‌دهد که در مقیاس زمانی روزانه، برخلاف روش نسبت نرمال، دقت مدل در بازسازی داده‌ها بیشتر است. چنان‌چه

پراکنش نقاط با نظم بهتری در اطراف خط یک‌به‌یک نمودارها قرار گرفته‌اند. ضریب همبستگی‌ها بیش از ۰/۶ است که البته با افزایش میزان نقص در آمار، به آرامی کاهش می‌یابد. ولی در کل با توجه به ضریب Nash دقت برآوردها چندان زیاد نیست. در مقیاس روزانه، مقادیر RMSE برآوردشده نسبت به روش نسبت نرمال نیز کمتر است که نشان‌دهنده دقت رویکرد همبستگی خطی نسبت به روش مذکور است. فاصله بین خط برازش داده شده بین داده‌های واقعی و برآورد شده در ابتدا کم است و با افزایش مقادیر بارش، این فاصله بیشتر می‌شود و این موضوع نشان‌دهنده آن است که دقت مدل در برآورد مقادیر حدی کمتر است. با توجه به اینکه خط برازش پایین‌تر از خط یک‌به‌یک قرار گرفته، روش رگرسیون خطی، مقادیر واقعی را کمتر از مقدار مشاهده‌ای برآورد می‌کند.

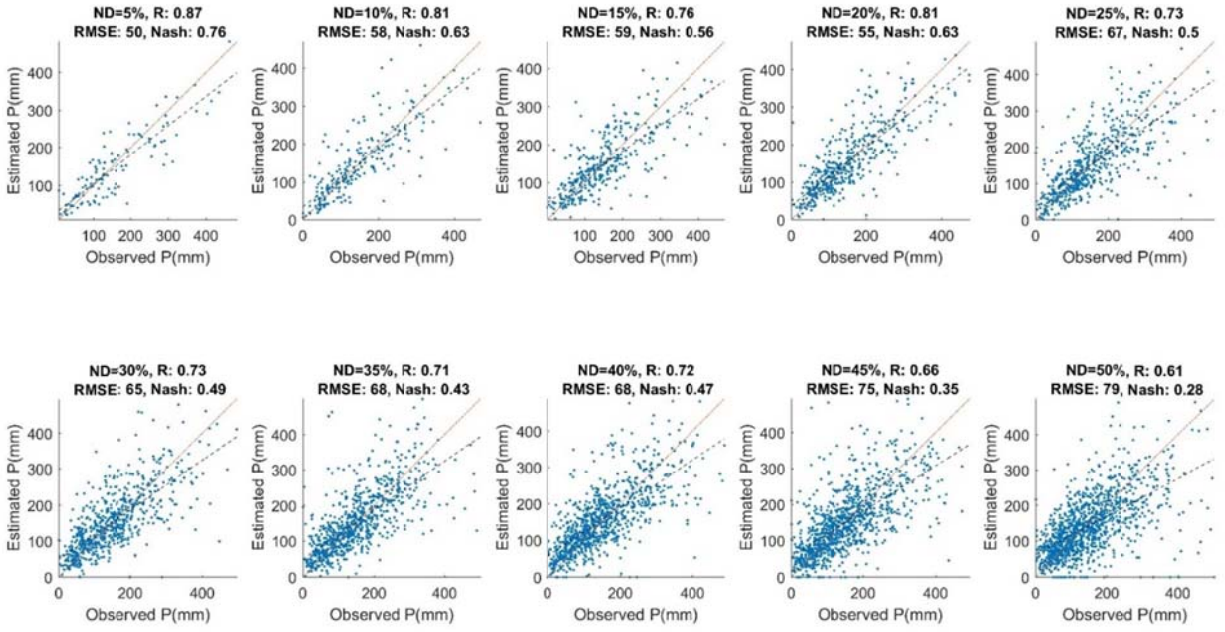
عملکرد روش رگرسیون خطی چند متغیره در شکل ۸، شکل ۹ و شکل ۱۰ به‌ترتیب در مقیاس زمانی روزانه، ماهانه و سالانه نمایش داده شده است. این شکل‌ها نشان می‌دهد که در مقیاس زمانی روزانه دقت این روش در بازسازی داده‌ها از رگرسیون خطی بیشتر است. برای مثال در هنگامی که نسبت نقص داده‌ها ۵ درصد است، شاخص همبستگی ۰/۷۶، مقدار RMSE ۱/۷ میلی‌متر و ضریب Nash ۰/۵۸ است که نسبت به روش‌های قبلی عملکرد بهتری را نشان می‌دهد. البته نکته قابل‌توجه در این روش، افت شدید شاخص‌های ارزیابی خصوصاً هنگامی است که نسبت نواقص آماری از ۳۰ درصد فراتر می‌رود. چنین حالتی برای بازسازی داده‌های ماهانه و سالانه هم وجود دارد و این روش به میزان نقص داده‌ها حساس است. شکل ۱۱، شکل۱۲ و شکل ۱۳ به‌ترتیب نشان‌دهنده پراکنش نقاط در رابطه با مقادیر بازسازی‌شده و واقعی بارش با استفاده از روش IDW و در مقیاس زمانی روزانه، ماهانه و سالانه است. در مقیاس زمانی روزانه، همانند روش‌های همبستگی خطی و رگرسیون چندمتغیره و برخلاف روش نسبت نرمال، داده‌های بازسازی‌شده و واقعی با پراکنش مشخصی اطراف خط یک‌به‌یک نمودارها ظاهر شده است. هر چند که با افزایش میزان نسبت نقص داده‌ها، این دقت کاهش می‌یابد.



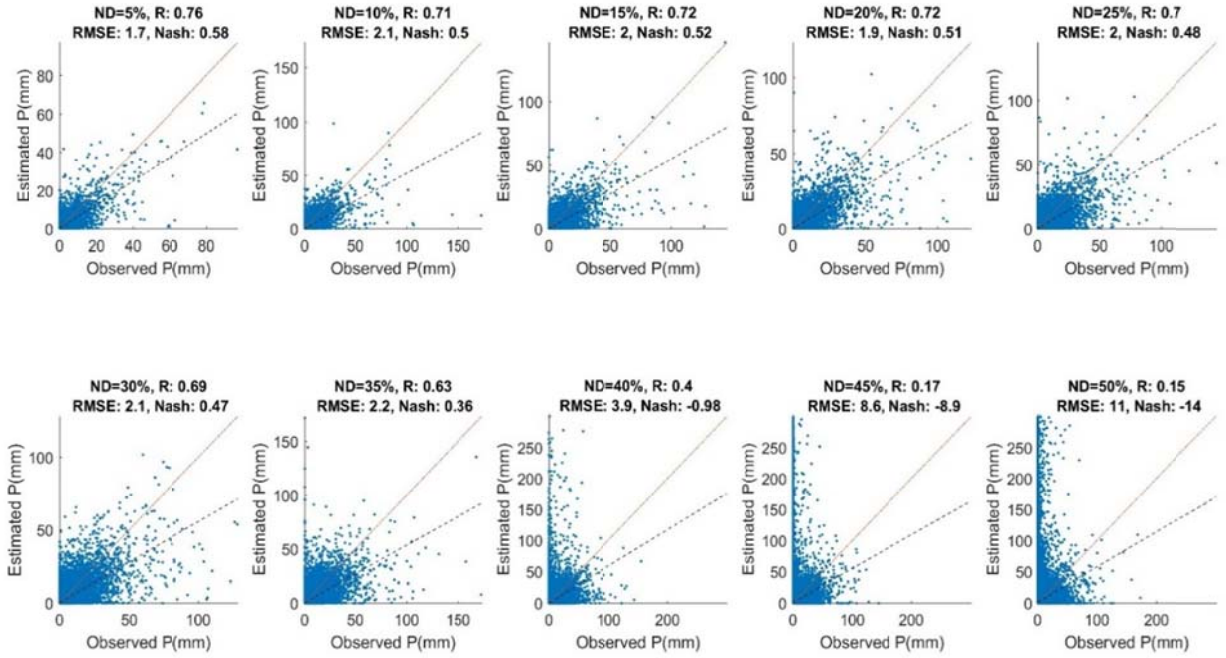
شکل۵. مقادیر بازسازی‌شده داده‌های بارش روزانه با استفاده از روش رگرسیون خطی در مقابل مقادیر واقعی بارش روزانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



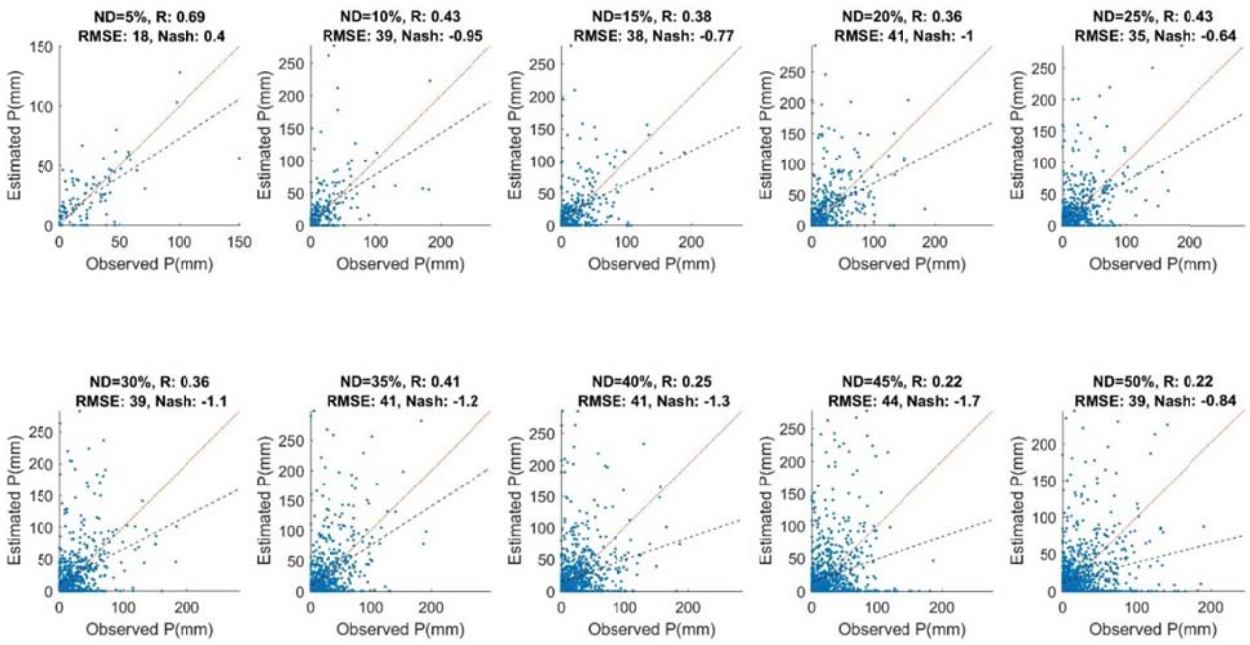
شکل۶. مقادیر بازسازی‌شده داده‌های بارش ماهانه با استفاده از روش رگرسیون خطی در مقابل مقادیر واقعی بارش ماهانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



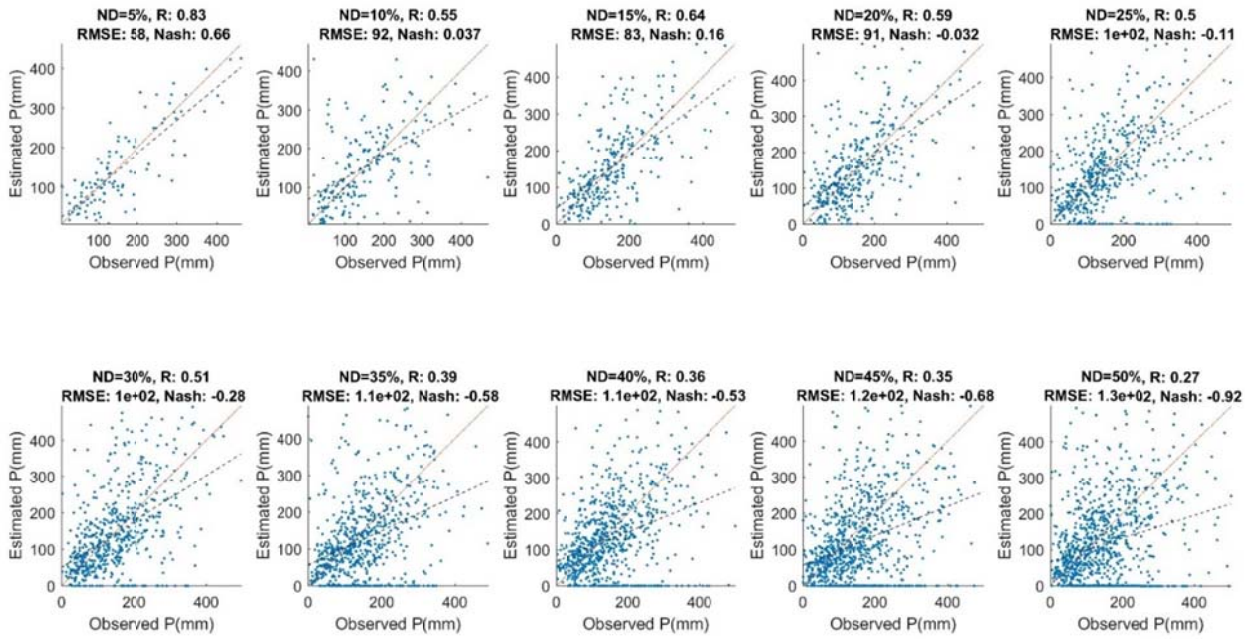
شکل۷. مقادیر بازسازی‌شده داده‌های بارش سالانه با استفاده از روش رگرسیون خطی در مقابل مقادیر واقعی بارش سالانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



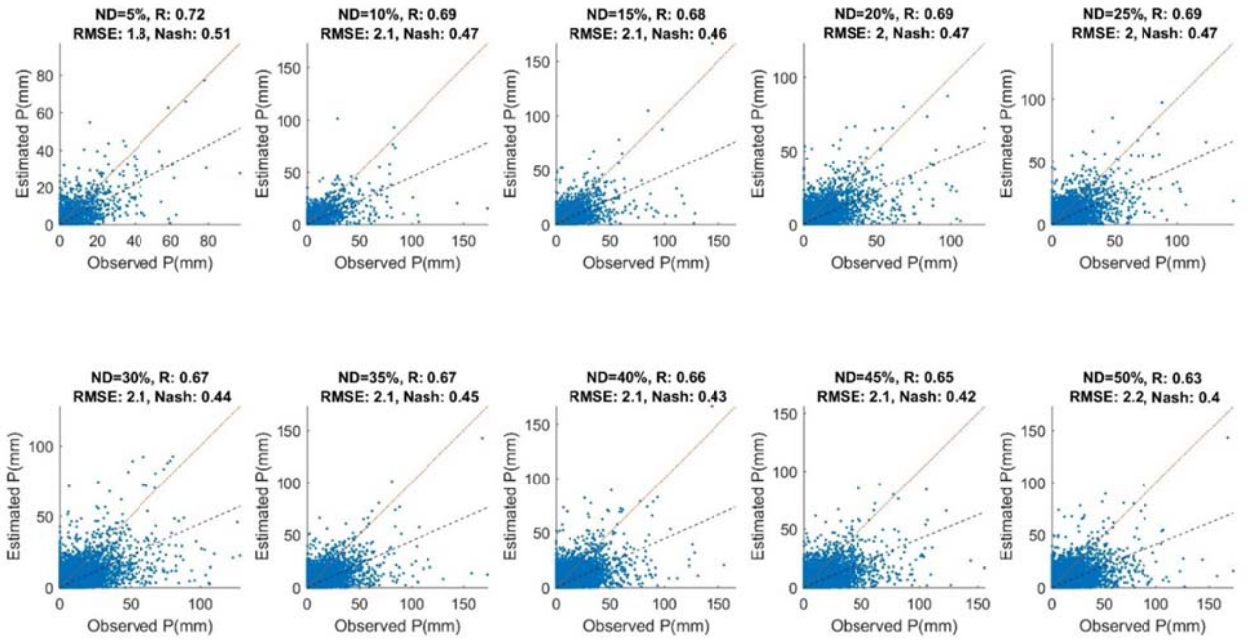
شکل۸. مقادیر بازسازی‌شده داده‌های بارش روزانه با استفاده از روش رگرسیون خطی چندمتغیره در مقابل مقادیر واقعی بارش روزانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



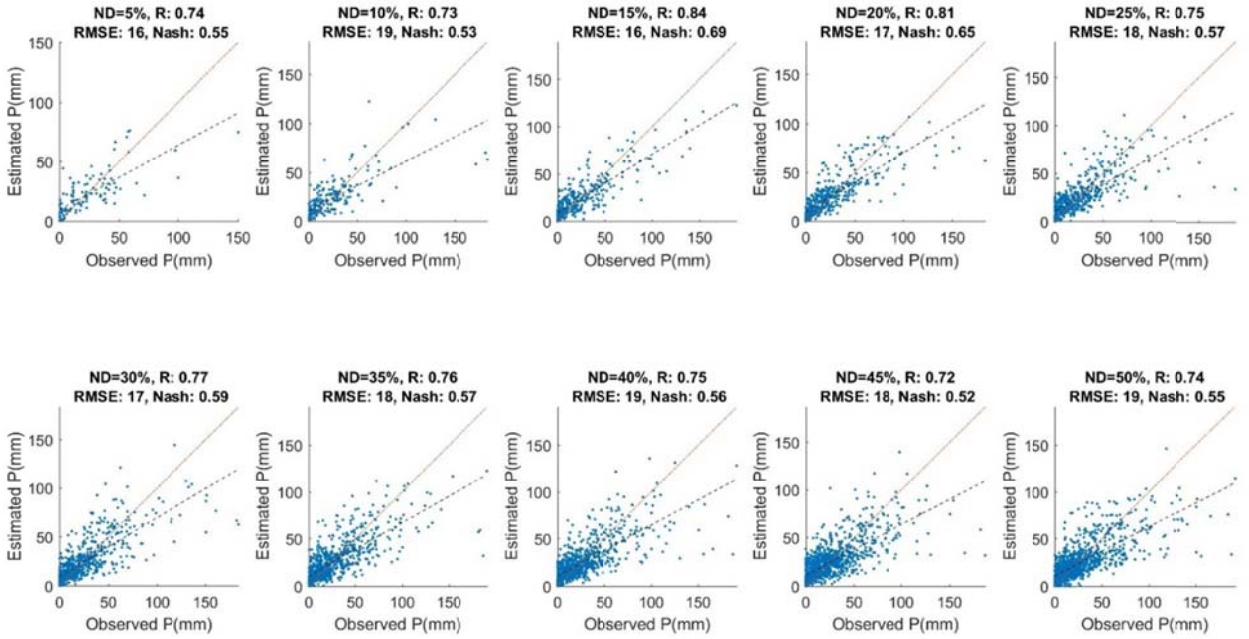
شکل۹. مقادیر بازسازی‌شده داده‌های بارش ماهانه با استفاده از روش رگرسیون خطی چندمتغیره در مقابل مقادیر واقعی بارش ماهانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



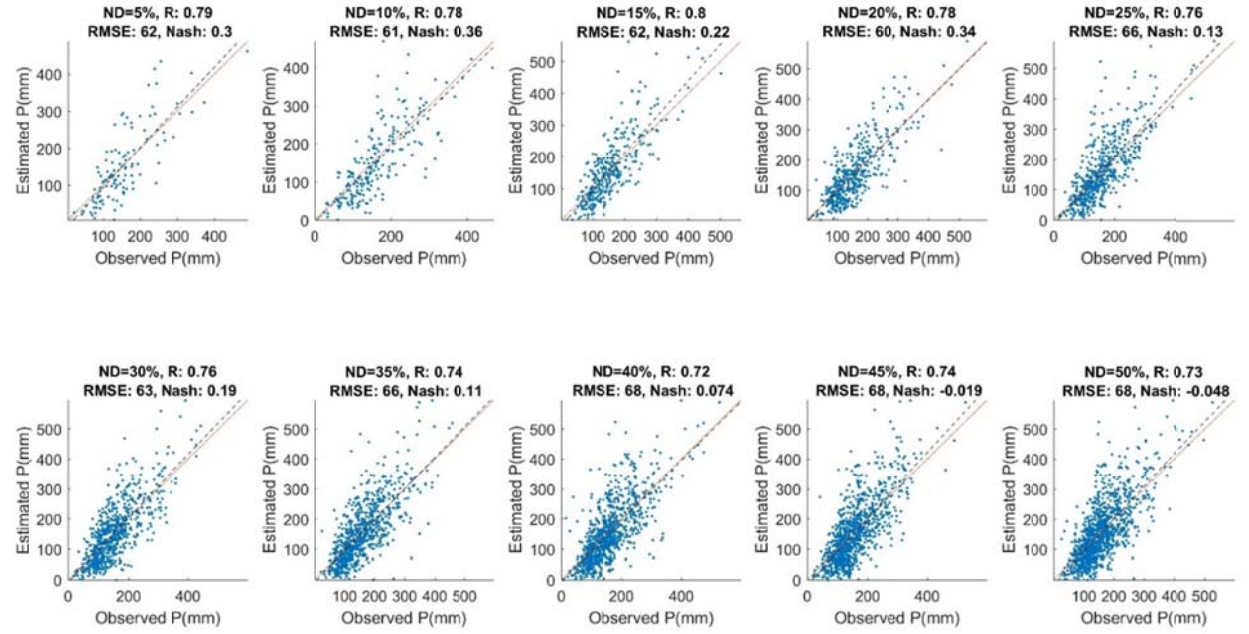
شکل ۱۰. مقادیر بازسازی‌شده داده‌های بارش سالانه با استفاده از روش رگرسیون خطی چندمتغیره در مقابل مقادیر واقعی بارش سالانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



شکل ۱۱. مقادیر بازسازی‌شده داده‌های بارش روزانه با استفاده از روش IDW در مقابل مقادیر واقعی بارش روزانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.



شکل۱۲. مقادیر بازسازی‌شده داده‌های بارش ماهانه با استفاده از IDW در مقابل مقادیر واقعی بارش ماهانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.

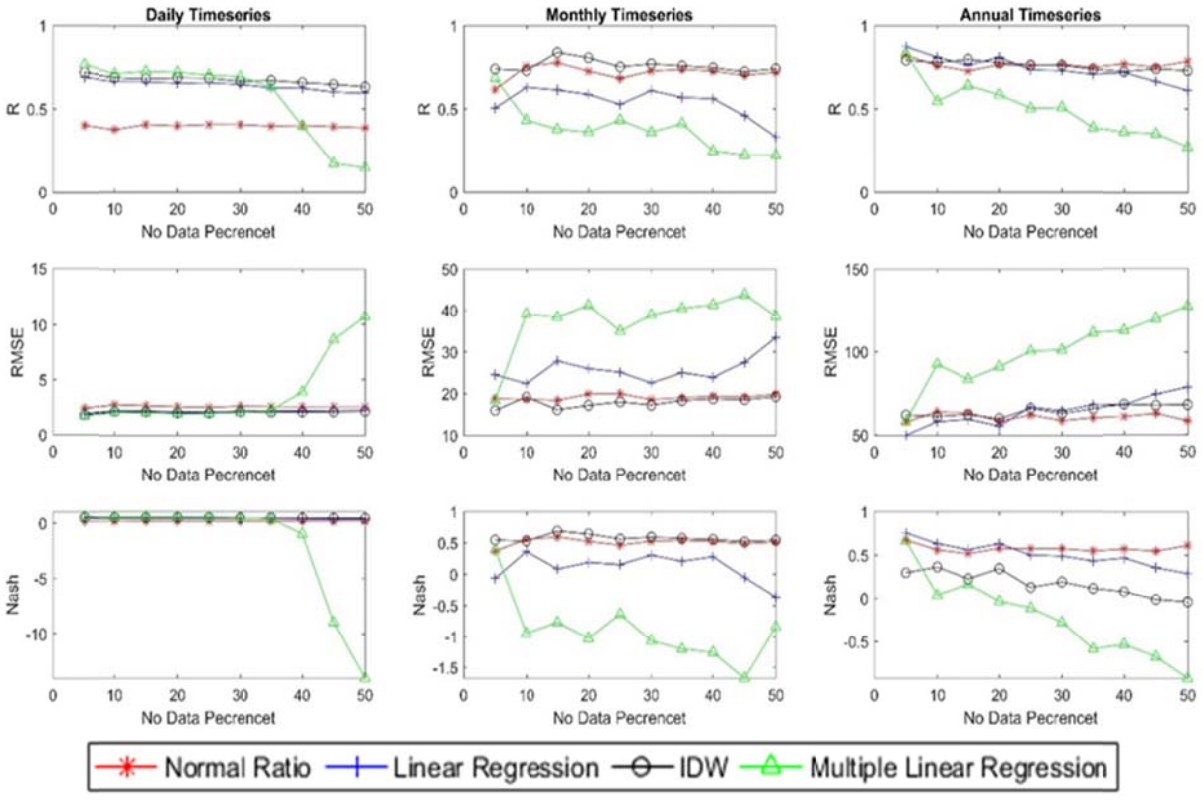


شکل۱۳. مقادیر بازسازی‌شده داده‌های بارش سالانه با استفاده از روش IDW در مقابل مقادیر واقعی بارش سالانه. مقادیر داده‌های ناقص به صورت ساختگی از ۵٪ تا ۵۰٪ داده‌ها ND در عنوان شکل‌ها به معنای No Data و درصد آن نیز نمایش داده شده است.

خلاصه‌ای از شاخص‌های ارزیابی روش‌ها را می‌توان در شکل ۱۴ مشاهده کرد که به‌خوبی عملکرد هر روش را نسبت به یکدیگر و همچنین افزایش میزان نسبت نقص داده‌ها نمایش می‌دهد. برای مثال اولین نمودار در این شکل حاکی‌از آن است که روش رگرسیون چندمتغیره خطی از لحاظ شاخص همبستگی R، RMSE و Nash تا هنگامی که نقص داده‌ها از ۳۰ درصد مقدار کل تجاوز نکرده است، از روش‌های دیگر بهتر است، در صورتی که

با افزایش نقص داده، به شدت از عملکرد آن کاسته می‌شود. این روش در سری‌های زمانی ماهانه و سالانه چندان مؤثر نیست و با افزایش اندک در نقص داده‌ها، از عملکرد آن به‌شدت کاسته می‌شود. در مقیاس روزانه، به‌ترتیب عملکرد روش رگرسیون چندمتغیره، IDW و همبستگی خطی بهتر است. البته در صورتی‌که نقص داده‌ها زیاد باشد، روش IDW و همبستگی خطی نتیجه بهتری دارند. در کل روش نسبت نرمال دقت کمتری نسبت به سایر روش‌ها دارد. در مقیاس ماهانه، روش IDW و نسبت نرمال دقت بیشتری نسبت به روش رگرسیون خطی و چندمتغیره دارد. با افزایش نقص داده‌ها، از عملکرد روش همبستگی خطی و چند متغیره، به‌شدت کاسته می‌شود. در مقیاس سالانه، هنگامی که نقص داده‌ها کمتر است، روش همبستگی خطی و سپس IDW دقت مناسبی دارند. در صورت افزایش نقص داده‌ها، روش نسبت نرمال با دقت بیشتری همراه است.

از لحاظ مقیاس زمانی، بیشترین دقت بازسازی بر اساس شاخص همبستگی R، در سری‌های سالانه است و بر مبنای همین شاخص، دقت بازسازی‌ها در مقیاس‌های سالانه بیشتر از ماهانه و در مقیاس‌های ماهانه بیشتر از روزانه است. در مقیاس زمانی سالانه و بر اساس شاخص R بیشترین دقت مربوط به روش همبستگی خطی برابر با ۰/۸۷ است. با استفاده از همین روش و نسبت به سایر مقیاس‌های زمانی، روش همبستگی خطی کمترین میزان شاخص RMSE برابر با ۵۰ میلی‌متر را نمایش می‌دهد. البته این میزان دقت در هنگامی است که میزان نقص در داده‌ها ۵ درصد باشد و با افزایش درصد نقص داده‌ها، از کارایی این روش کاسته می‌شود. این در حالی است که روش نسبت نرمال در مقابل افزایش درصد نقص داده‌ها مقاومت بیشتری نشان می‌دهد. در مرتبه بعدی، مقاومت روش IDW نسبت به افزایش درصد داده‌های گم‌شده بیشتر است.



شکل ۱۴. تغییرات شاخص‌های R، RMSE و Nash با تغییر در نسبت داده‌های گم‌شده و در رابطه با روش‌های مختلف بازسازی داده و مقیاس‌های زمانی روزانه، ماهانه و سالانه.

۵. بحث و نتیجه‌گیری

در بازسازی نواقص آماری داده‌های بارش، نسبت به مقیاس زمانی داده و میزان نقص در داده‌ها، هر روش عملکرد متفاوتی دارد. در کل روش IDW و نسبت نرمال از پایداری بیشتری نسبت به روش همبستگی خطی و خصوصاً رگرسیون چندمتغیره برخوردار هستند. نتایج نشان داد که در بازسازی نواقص آماری بارش روزانه، روش نسبت نرمال دقت بسیار کمتری نسبت به IDW یا رگرسیون خطی و چندمتغیره دارد. متناسب با رابطه محاسبه نواقص آماری بر اساس روش نسبت نرمال، آنچه در وزنِ تأثیرگذاری ایستگاه‌های همجوار در بازسازی نقص آماری تأثیر دارد، مقدار تشابه میانگین سری‌های زمانی است. بر خلاف روش IDW که نزدیکی فاصله در تأثیرگذاری ایستگاه موثر است و یا روش رگرسیون خطی و چندمتغیره که میزان همبستگی بین سری‌های زمانی در تعیین ایستگاه منتخب برای رفع نواقص آماری ملاک قرار می‌گیرد. در سری‌های زمانی بارش روزانه در مناطق خشک کشور، به خاطر آن‌که بخش قابل‌توجهی از روزهای سال بدون بارش است، لذا میانگین بیشتر ایستگاه‌ها نزدیک به یکدیگر می‌باشند و اهمیت و نقش ایستگاه‌های مختلف تقریباً یکسان است. در نتیجه، بازسازی نواقص آماری سری‌های زمانی بارش روزانه به روش نسبت نرمال دقت مناسبی ندارد. این در حالی است که روش‌های IDW و رگرسیون خطی به مراتب مناسب‌تر عمل می‌کنند. هنگامی که سری زمانی داده‌ها به‌صورت ماهانه و سالانه مورد استفاده قرار می‌گیرند، دقت روش نسبت نرمال نیز افزایش قابل‌توجهی می‌یابد. این موضوع به خاطر نزدیک شدن میانگین‌های بارش ایستگاه‌های مشابه در سری‌های زمانی ماهانه است. لذا ایستگاه‌های مشابه از لحاظ میانگین بارش، نقش خود را در بازسازی نواقص آماری بهتر ایفا می‌کنند.

در مقیاس سالانه و در حداقل مقدار ممکن نقص در داده‌های اولیه ۵ درصد و بر اساس روش همبستگی خطی، خطای محاسبات به‌طور متوسط حدود ۵۰ میلی‌متر می‌باشد. هنگامی که نقص داده‌ها افزایش می‌یابد و بر اساس روش نسبت نرمال، این میزان به حدود ۶۰ میلی‌متر می‌رسد. هر چند که در اینجا از روش‌های مختلفی استفاده شد و قابلیت آنها موردبررسی قرارگرفت، باید در نظر داشت که در مناطق خشک ایران که گاهاً متوسط بارندگی برخی ایستگاه‌ها کمتر از ۱۰۰ میلی‌متر می‌باشد، این میزان خطا قابل توجه است. البته در اینجا خطای متوسط مربوط به کل ایستگاه‌های موردمطالعه بوده که متوسط بارش در ایستگاه‌های مورد بررسی در حدود ۱۵۰ میلی‌متر است. یعنی به‌طور متوسط، به‌میزان یک‌سوم از میانگین، احتمال خطا وجود دارد. هر چند که نقاط بازسازی شده در مقابل مقادیر واقعی پراکنش مناسبی را در اطراف خط مرکزی نمودار نشان می‌دهند، به هر حال خطایی در حدود ۶۰ میلی‌متر در بازسازی نواقص آماری سالانه انتظار است. البته باید در نظر داشت که پراکنش زمانی و مکانی بارش در مناطق خشک بسیار زیاد است و همین موضوع، تخمین نواقص آماری را با چالش همراه می‌سازد.

یافته‌های این پژوهش نشان می‌دهد که درکل دقت بازسازی‌ها در مقیاس‌های سالانه بیشتر از ماهانه و در مقیاس‌های ماهانه بیشتر از روزانه است. این موضوع به خاطر کاهش تضاریس داده‌ها و هموارتر شدن آن‌ها در سری‌های زمانی ماهانه و سالانه نسبت به روزانه است. در واقع هر چه این میزان تضاریس کاهش یابد، دقت روش‌ها در بازسازی نواقص افزایش می‌یابد. در مقیاس روزانه، دقت روش‌های بازسازی IDW با اندکی تفاوت بیشتر از روش همبستگی خطی است که فاصله بسیار زیادی با روش نسبت نرمال دارند. البته دقت روش IDW نیز کامل نیست. در واقع دو ایستگاه نسبتاً نزدیک در صورت قرار گرفتن در قسمت‌های مخالف یک رشته‌کوه می‌توانند تفاوت‌های اساسی در میانگین آب‌وهوا و تغییرپذیری آب‌وهوا داشته باشند (باریوس و همکاران، ۲۰۱۸).

به‌هر حال هر روش محاسن و معایبی دارد. استفاده از روش‌های پیچیده‌تر مانند درون‌یابی و تخمین داده‌های

ناقص با استفاده از کریجینگ، هرچند که ممکن است با دقت بیشتری همراه شود، به‌غیر مباحث مربوط به تنظیمات اولیه قابل‌توجه کریجینگ و محاسبات سنگین آن، شرایط مربوط به نرمال بودن داده‌های ووردی نیز بایستی مد نظر قرار گیرد. برای مثال نتایج پژوهشی آهنی و همکاران (۲۰۱۲) نشان داد که برای درون‌یابی و تخمین حجم بارش در استان فارس، کریجینگ با محدودیت عدم‌نرمال بودن توزیع فراوانی بارش در بسیاری از ماه‌های سال همراه است. لذا از روش IDW به‌عنوان روش جایگزین استفاده شد. از دیگر محدودیت‌های روش کریجینگ مشکلات تعیین semi-variogram مناسب برای داده‌ها، عدم‌قطعیت‌های مربوط به تعیین نقاط sill و nugget است (تیگاورپو، ۲۰۲۰). همچنین نتایج تحقیق دیگری (اشتیلی‌نوا و همکاران، ۲۰۱۷) نشان داد که تکنیک‌های با محاسبات زیاد مانند کریجینگ ممکن است تأثیر چندانی در بهبود دقت بازسازی‌ها ایجاد نکند. علاوه‌بر آن شاید استفاده از روش‌های هوش مصنوعی در بازسازی داده‌ها با دقت بیشتری همراه شود، ولی به‌هر حال نیاز به تولید تعداد فراوان مدل (حسن‌پورکاشانی و دینپشوه، ۲۰۱۲) خصوصاً در شرایطی از که داده‌های روزانه استفاده می‌شود، حجم عملیات بازسازی را به‌شدت زمان بر می‌کند.

مقیاس مطالعات فعلی در سطح کشور ایران است. چنان‌چه اطلاعات مربوط به ایستگاه‌های باران‌سنجی ذخیره‌ای مربوط به سازمان هواشناسی و وزارت نیرو نیز به این داده‌ها اضافه شوند، انتظار می‌رود که دقت عملکرد روش‌ها افزایش یابد. همان‌طور که نتایج تحقیق حاضر نشان داد، با افزایش نسبت داده‌های ناقص، دقت مدل‌ها کاهش می‌یابد. لذا چنان‌چه داده‌های جدیدی وارد محاسبات شوند و در واقع تراکم ایستگاه‌های باران‌سنجی افزایش یابد، انتظار عملکرد بهتر هر یک از این روش‌ها نیز وجود دارد. همچنین نتایج (لبرنز و همکاران، ۲۰۱۶) نشان داده است که دقت بازسازی نواقص آماری داده‌های بارش به تراکم و تعداد ایستگاه‌های باران‌سنجی بستگی دارد.

شاید بتوان گفت که مهم‌ترین یافته این پژوهش، تفاوت عملکرد روش‌های مورداستفاده متناسب با نسبت نواقص آماری موجود در داده‌ها و همچنین نوع مقیاس سری زمانی است. در واقع استفاده از هر روش، به شرایط موجود بستگی دارد. نوع سری زمانی و پارامتر موردنظر، فصل مقادیر گمشده، میزان مقادیر گم‌شده در داده‌ها، منطقه آب‌وهوایی، دانش و تخصص شخص مسئول تصحیح داده‌ها، امکانات در دسترس خصوصاً از لحاظ سخت‌افزاری و نرم‌افزاری، توانایی و دقت روش، طول ضبط داده‌های موجود و ... همگی بر روی عملکرد نهایی روش بازسازی داده‌های ناقص تأثیرگذارند. بدین‌جهت می‌توان گفت که فرایند بازسازی داده‌ها یک فرایند پویا است. هر چند نتایج حاصل از این مطالعه می‌تواند به‌عنوان یک راهنمای پایه، در اختیار محققان و کارشناسان مرتبط با داده‌های هواشناسی به‌خصوص اقلیم‌شناسان و هیدرولوژیست‌ها قرار گیرد، ولی متناسب با شرایط بایستی تصمیم گرفت که از چه روشی استفاده شود و حتماً عملکرد آن نیز موردارزیابی قرار گیرد. لذا لازم است تا فرایند بازسازی داده‌ها خصوصاً در مطالعات علوم محیطی به‌صورت یک نرم‌افزار واحد ارائه شود تا هم فرایندهای بازسازی داده یک‌پارچه شود و هم علاوه‌بر کوتاه‌ترشدن این عملیات وقت‌گیر از لحاظ حجم محاسبات، موردارزیابی نیز قرار گرفته و در نهایت روشی انتخاب شود که متناسب با شرایط بهترین عملکرد را نسبت به سایرین داراست. علاوه‌بر آن می‌توان از این نرم‌افزار، همانند بسته نرم‌افزاری که توسط سرانو_نوتیویلی و همکاران (۲۰۱۷) و یا هو و هوانگ (۲۰۲۰) در محیط نرم‌افزار R برای تولید داده‌های بارش روزانه ایجاد شده است، جهت تولید نقشه‌های پارامترهای هواشناسی در نقاط فاقد داده نیز استفاده کرد. به‌هر حال در قالب یک بسته نرم‌افزاری بهتر است دقت سایر روش‌های بازسازی نواقص آماری خصوصاً روش‌های مبتنی بر هوش

مصنوعی نیز مورد بررسی قرار گیرند که در تحقیقات آتی بدان پرداخته خواهد شد.

مراجع

رضازاده جودی.، ع. و ستاری، م. ت.، ۱۳۹۵، ارزیابی عملکرد روش‌های مختلف در بازسازی داده‌های بارش. تحقیقات کاربردی علوم جغرافیایی، (۱۶)۴۲، ۱۷۶-۱۵۵.

ساداتی نژاد، ج.، ۱۳۷۶، مقایسه آماری و روش‌های مختلف بازسازی داده‌های بارش در استان اصفهان. پایان نامه ارشد، دانشگاه تربیت مدرس.

علیزاده، ا.، ۱۳۹۲، اصول هیدرولوژی کاربردی. چاپ ۳۶، دانشگاه امام رضا.

مهدوی، م.، ۱۳۸۴، هیدرلوژی کاربردی. جلد اول، انتشارات دانشگاه تهران.

رضیئی، ط.، ۱۳۹۶، چشم‌اندازی از مناطق اقلیمی ایران به روش کوپن-گایگر در سده بیست و یکم، م. فیزیک زمین و فضا، (۲)۴۳، ۴۱۹-۴۳۹.

Abebe, A. J., Solomatine, D. P. and Venneker, R. G. W., 2000, Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrological Sciences Journal, 45, 425–436.

Ahani, H., Kherad, M., Kousari, M., Rezaeian-Zadeh, M., Karampour, M., Ejraee, F. and Kamali, S., 2012, An investigation of trends in precipitation volume for the last three decades in different regions of Fars province, Iran. Theoretical and Applied Climatology, 109, 361-382. doi: 10.1007/s00704-0-572-11z.

Barlow, M., Zaitchik, B., Paz, S., Black, E., Evans, J. and Hoell, A., 2016, A Review of Drought in the Middle East and Southwest Asia. Journal of Climate, 29, 8547-8574. doi: 10.1175/jcli-d-13-00692.1.

Barrios, A., Trincado, G. and Garreaud, R., 2018, Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. Forest Ecosystems, 5, 28. doi: 10.1186/s40663-018-0147-x.

Canchala-Nastar, T., Carvajal-Escobar, Y., Alfonso-Morales, W., Loaiza Cerón, W. and Caicedo, E., 2019, Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural network Data in Brief 26, 104517. doi: https://doi.org/10.1016/j.dib.2019.104517.

De Martonne, E., 1925, Traité de Géographie Physique Quatrième édition. Paris: A. Colin.

Foehn, A., García Hernández, J., Schaefli, B. and Cesare, G., 2018, Spatial interpolation of precipitation from multiple rain gauge networks and weather radar data for operational applications in Alpine catchments. Journal of Hydrology, 563, 1092-1110. doi: https://doi.org/10.1016/j.jhydrol.2018.05.027.

Hasanpour Kashani, M. and Dinpashoh, Y., 2012, Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment, 26, 59-71. doi: 10.1007/s00477-011-0536-y.

Hu, M. and Huang, Y., 2020, atakrig: An R package for multivariate area-to-area and area-to-point kriging predictions. Computers & Geosciences, 139, 104471. doi: https://doi.org/10.1016/j.cageo.2020.104471.

Kamwaga, S., Mulungu, D. M. M. and Valimba, P., 2018, Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania. Physics and Chemistry of the Earth, Parts A/B/C , 106, 17-28. doi: https://doi.org/10.1016/j.pce.2018.05.008.

Kim, J.-W. and Pachepsky, Y. A., 2010, Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. Journal of Hydrology, 394, 305-314. doi: https://doi.org/10.1016/j.jhydrol.2010.09.005.

Lebrenz, H., Bárdossy, A. and Pavia Santolamazza, D., 2016, Reconstruction of missing precipitation data. Paper presented at the EGU General Assembly, Vienna.

Lo Presti, R., Barca, E. and Passarella, G., 2008, A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environmental Monitoring and Assessment, 160, doi: 10.1007/s10661-008-0653-3.

Miri, M., Masoudi, R. and Raziei, T., 2019, Performance Evaluation of Three Satellites-Based Precipitation Data Sets Over Iran, Journal of the Indian Society of Remote Sensing, 47, pages2073–2084.

Rees, G., 2008, Hydrological Data. In: Gustard, Alan; Demuth, Siegfried, (eds.). Manual on Low-flow Estimation and Prediction Operational Hydrology Report, World Meteorological Organization,50, 22-35.

Sattari, M.-T., Rezazadeh-Joudi, A. and Kusiak, A., 2016, Assessment of different methods for estimation of missing data in precipitation

studies. Hydrology Research, 48(4), 1032-1044. doi: 10.2166/nh.2016.364.

Serrano-Notivoli, R., de Luis, M. and Beguería, S., 2017, An R package for daily precipitation climate series reconstruction. Environmental Modelling & Software, 89, 190-195. doi: https://doi.org/10.1016/j.envsoft.2016.11.005.

Shtiliyanova, A., Bellocchi, G., Borras, D., Eza, U., Martin, R. and Carrère, P., 2017, Kriging-based approach to predict missing air temperature data. Computers and Electronics in Agriculture, 142, 440-449. doi: https://doi.org/10.1016/j.compag.2017.09.033.

Teegavarapu, R. S. V., 2020, Precipitation imputation with probability space-based weighting methods. Journal of Hydrology, 581, 124447. doi: https://doi.org/10.1016/j.jhydrol.2019.124447.

Teegavarapu, R. S. V., Aly, A., Pathak, C. S., Ahlquist, J., Fuelberg, H. and Hood, J., 2018, Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections. International Journal of Climatology, 38, 776-793. doi: 10.1002/joc.5209.

Journal of the Earth and Space Physics, Vol. 47, No. 2, Summer 2021, P. 8

Investigation of the efficiency of methods of infilling missing data in relation to the precipitation parameter in arid regions of Iran

Kousari, M. R.[1*], Esmaeilzadeh Hosseini, M.[2] and Miri, M.[1]

1. Assistant Professor, Soil Conservation and Watershed Management Research Institute (SCWMRI), Agricultural Research, Education and Extension Organization (AREEO), Tehran, Iran
2. Ph.D. Student, Department of Arid Land And Desert Management, Faculty of Natural Resources & Desert Studies, Yazd University, Yazd, Iran

(Received: 5 Feb 2021, Accepted: 25 May 2021)

Summary

Missing data are common issue in climate data. Also precipitation is a very important part of the hydrological cycle and meteorological and hydrological studies of watersheds, initially depend on the quantity and quality of recorded rainfall data and its distribution in the area. Complete and reliable sets of climatic and hydrological data are required for planning and design of these projects. Therefore for treatment of precipitation missing data, various methods have been developed and applied. Normal ratio method, linear regression, multivariate regression and inverse distance weighting (IDW) have a wide applications in natural resources study in our country. Therefore, it is necessary to determine the ability of these methods, especially in relation to the precipitation parameter, which plays a crucial role in the study of natural resources. In this study, the capability of each mentioned methods for infilling missing data of daily, monthly and annual precipitation time series in the arid regions of Iran was investigated for varying proportion of missing data from 5 to 50% of total data. In fact, the main purpose of this study is to answer the question of which of the four mentioned methods are more effective for infilling precipitation missing data.

The daily data of Iran's synoptic meteorological stations were used for the present study. Using the Run homogeneity test, the data homogeneity was investigated. Also, using graphical exploring data, and especially boxplot diagrams, outlier data were identified and flagged as missing data. The average annual precipitation and temperature of 400 stations were determined, and then based on these data their de Martonne coefficients were computed. In the next step, stations with de Martonne coefficient less than 10 were selected as arid climate. Among them, 73 stations that had sufficient data from 1986 to 2017 were distinguished. To evaluate each of the data reconstruction methods, part of the actual data was deliberately discarded from the original data and then reconstructed. Due to high volume of calculations, this process was programmed in MATLAB software.

The results showed that each method had different functionality according to the conditions. Daily data are not well estimated using the normal ratio method to estimate the missing data less than the actual one. The use of linear regression method showed that in daily time scale, unlike the normal ratio method, the model accuracy in data reconstruction is higher. For linear regression approach, the distance between the fitted line between the observed and estimated data is small at first, and as the precipitation increases, this distance increases, indicating that the model is less accurate in estimating the extreme values. Given that the fitting line is below the 1:1 line, the linear regression method estimates the actual values below normal. The same results can be found for IDW producer. The multivariate regression method is more accurate for daily time series when the proportion of missing data are not considerable, but is generally very sensitive to the proportion of missing data. The normal ratio method is not suitable for reconstructing daily missing values, however it is more stable than other methods when missing data increase. In monthly time series, the performance of the IDW method and then the normal ratio is better. In annual series, linear correlation, normal ratio, and IDW have better performances, respectively.

The findings of this study show that in general, the accuracy of reconstructions on annual scales is more than monthly and on monthly scales is higher than daily. This is due to smoother time series in the monthly and annual time series than the daily ones. Also it should be noted that the scale of current studies is in Iran. If the data from the reserved rain-gauge stations of the Meteorological Organization and the Ministry of Energy are added to this data, the accuracy of the methods is expected to increase. As the results of the present study show, the accuracy of the models decreases with increasing incomplete data ratio. Therefore, if new data is included in missing data processing, there is an expectation of better performance of each of these methods. Finally it should be considered that each method should be used in accordance with the given conditions, and therefore it is recommended to develop a software package for infilling missing data in Iran.

Keywords: Gap in data, linear regression, Normal Ratio, infilling, precipitation.

* Corresponding author: mohammad_kousari@yahoo.com