تخمین داده‌های گمشده بارش و رواناب روزانه با استفاده از نگاشت خودسامانده (مطالعه موردی استان مازندران)

ساناز اسلامی جمال آباد[۱]، احمد شرافتی[۱*]، عمادالدین محمدی گل افشانی[۱] و فرهاد فرسادنیا[۲]

(تاریخ دریافت: ۱۳۹۶/۱۰/۳ ؛ تاریخ پذیرش: ۱۳۹۷/۳/۶)

چکیده

یکی از مشکلات پیش‌روی متخصصان و طراحان پروژه‌های آبی، سری‌های زمانی ناقص در علم هیدرولوژی است که باعث بروز خطا در نتایج مطالعات شده و اجرای پروژه‌ها را دچار مشکل می‌کند. این مساله در مناطقی که تعداد ایستگاه‌های هیدرومتری و باران‌سنجی محدود است، حادتر است. از طرفی ارزیابی، توسعه و استفاده پایدار از منابع آبی نیازمند در اختیار داشتن سری‌های زمانی هیدرولوژیکی با کیفیت بالا و طول مدت کافی است. لذا این موضوع رفع نواقص آماری را ایجاب می‌کند و اهمیت چگونگی مواجه شدن با این مشکلات در آنالیزهای هیدرولوژیکی را نشان می‌دهد. در حال حاضر استفاده از روش‌های آماری به‌منظور رفع خلاءهای آماری و بازسازی داده‌ها متداول است. در این مطالعه به‌منظور معرفی روشی چندمتغیره در برآورد داده‌های گمشده مربوط به بارش و رواناب، در یک منطقه همگن از لحاظ هیدرولوژیکی در استان مازندران، روش نگاشت‌های خودسازمانده تحت دو سناریو مورد بررسی قرار گرفت و تخمین‌های قابل اعتمادی را به‌دست آورد. به‌نحوی که مقادیر ضریب همبستگی بین داده‌های مشاهداتی و خروجی مدل برای داده‌های بارش تا ۰/۹۲ و برای داده‌های رواناب تا ۰/۹۵ محاسبه شد. لذا پیشنهاد می‌شود برای کاهش عدم اطمینان ناشی از داده‌های ناکافی در مدیریت منابع آب، از این روش استفاده شود.

واژه‌های کلیدی: شبکه‌عصبی غیرنظارت شونده، بازسازی سری‌های زمانی، داده‌های گمشده

۱. گروه مدیریت ساخت و آب، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی تهران، ایران
۲. گروه مهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد
*: مسئول مکاتبات: پست الکترونیکی: aharafati@srbiau.ac.ir

مقدمه

اساس و پایه مطالعات هیدرولوژی، داده‌های آماری مورد قبول است. با توجه به خلأهای گسسته و پیوسته در اغلب داده‌های هیدرولوژی مانند دبی رودخانه‌ها که به‌دلایلی مانند عدم ثبت آمار، حذف آمار غلط و خرابی یا از بین رفتن دستگاه‌های اندازه‌گیری اتفاق می‌افتد، تخمین و برآورد این داده‌ها لازم و ضروری است. دسترسی به داده‌های کافی و دقیق از یک‌سو موجب کوتاه‌تر شدن مدت مطالعات و ازسوی دیگر موجب براورد دقیق‌تر پارامترهای هدف و کاهش هزینه‌های اجرایی و خسارت بعدی ناشی از اجرای طرح‌ها می‌شود (۲۲). با این وجود اکثر داده‌های هیدرولوژیکی که در کشور یافت می‌شود، به‌دلایل مختلف از جمله وقوع جنگ تحمیلی و عدم اندازه‌گیری، تغییر سیاست‌های نهادهای مسئول، خرابی دستگاه‌های اندازه‌گیری و یا سهل‌انگاری کاربران، از نظر طول دوره ناکافی هستند یا در کیفیت آنها شک و تردید وجود دارد و یا در طول دوره داده‌برداری وجود داده‌های مفقود شده، مرسوم و رایج است. موضوع ذکر شده خاص ایران نیست و این شرایط در بیشتر کشورهای در حال توسعه مشاهده می‌شود (۵، ۱۳، ۱۵). در ایجاد این کمبود داده‌ها عوامل مختلفی نقش دارند، مانند غیبت موقت مسئول داده‌برداری، خرابی وسایل نظارتی و یا کمبود منابع مالی و غیره. نتیجه استفاده از این داده‌ها عدم اطمینان و درنهایت عملکرد ضعیف سیستم‌های مدیریت منابع آب خواهد بود (۳ و ۵).

برخی روش‌های تخمین داده‌های گم شده به تعداد داده‌های گم شده، در دسترس بودن داده‌های هواشناسی ایستگاه‌های اطراف، فصلی از سال که داده‌های گم شده در آن واقع هستند، نوع اقلیم منطقه، سطح دانش و مهارت جمع آورنده اطلاعات، طول مدت داده برداری، اهمیت دقت در پیش‌بینی وابسته هستند (۱۳، ۱۸). این روش‌ها از یک درون‌یابی ساده شروع شده و به روش‌های آنالیز آماری پیچیده ختم می‌شوند. هنگام مواجهه با یک auto- series (داده‌های یک ایستگاه که برای پر کردن داده گمشده استفاده می‌شود) معمولاً روش‌هایی از قبیل میانگین‌گیری حسابی مورد استفاده قرار می‌گیرد (۱۰ و ۲۳) و یا روش‌های درون‌یابی خطی استفاده می‌شود. درون‌یابی خطی شامل کشیدن یک خط راست بین دو نقطه می‌شود. این نقاط در دو طرف مکانی هستند که در آنجا داده وجود ندارد و درون‌یابی توسط این خط مستقیم صورت می‌گیرد.

معمولاً استفاده از داده‌های ایستگاه‌های مجاور برای محاسبه میانگین وزنی مرسوم‌تر است (۱۳ و ۲۱). همچنین برای محاسبه رگرسیون خطی نیز از داده های ایستگاه های مجاور استفاده می شود (۲ و ۹). فاکتور وزن ممکن است از ضریب همبستگی یا نسبت مساحت‌ها یا فاصله بین ایستگاه بدون داده و ایستگاه‌های داده دهنده برآورد شود (۶). یو و همکاران در پژوهشی از دو روش عکس مجذور فاصله (گرافیکی) و رگرسیون برای تخمین دما استفاده کردند و این دو روش را ارزیابی کردند و مقادیر تخمین زده شده را با مقادیر اندازه‌گیری شده واقعی مقایسه کردند. آنها ضریب کارایی، واریانس، ریشه میانگین مربعات خطا و خطاهای سیستماتیک و غیرسیستماتیک را محاسبه کردند و به این نتیجه رسیدند که روش رگرسیون در نواحی کوهستانی مناسب‌تر از عکس مجذور فاصله است و نیز در مواردی که تعداد ایستگاه‌های هواشناسی پراکنش کمی داشته باشند هر دو روش ناتوان هستند. عبداللهی (۱) با مقایسه الگوریتم لاگرانژ و روش‌های متداول بازسازی داده‌های هیدرولیکی، نتیجه گرفت که الگوریتم فراکاوشی لاگرانژ کمترین مقدار ریشه میانگین مربعات خطا را نسبت به روش‌های ایستگاه معرف، نسبت نرمال و روش سازمان جهانی هواشناسی دارا است و مقادیر بازسازی شده به مقادیر مشاهداتی نزدیک‌تر است.

اگر چه بیشتر روش‌های سنتی ساده‌سازی را پیشنهاد می‌کنند ولی روش‌های پیچیده‌تری نیز وجود دارد. به‌عنوان مثال، ساداتی‌نژاد (۲۹) در مقایسه آماری روش‌های مختلف بازسازی داده‌های بارش سالانه در استان اصفهان به‌منظور تعیین بهترین روش در هر اقلیم، هفت روش بازسازی شامل ایستگاه معرف، نسبت نرمال، محور مختصاتی، رگرسیون خطی ساده، رگرسیون

خطی چندمتغیره و اتورگرسیو را مورد بررسی قرار داد که در این مقایسه روش نسبت نرمال را برای اقلیم‌های خشک و مدیترانه‌ای و روش رگرسیون خطی چندمتغیره را برای اقلیم نیمه‌خشک به‌عنوان روش‌های برتر معرفی کرد. همچنین ساداتی‌نژاد و همکاران (۳۰) در پژوهشی کارایی روش رگرسیون خطی فازی در مقایسه با روش‌های رگرسیون خطی ساده، رگرسیون چندمتغیره و روش محور مختصاتی و نسبت نرمال در مورد بازسازی دبی سالانه ایستگاه‌های هیدرومتری حوضه آبریز کارون بزرگ را مورد ارزیابی قرار دادند و درنهایت روش رگرسیون خطی ساده را به‌عنوان روش برتر در بازسازی داده‌های دبی سالانه معرفی کردند. جنبه مهم دیگر بعضی از این روش‌های سنتی، فرض ضمنی یا صریح بودن رابطه خطی بین متغیرها است که ممکن است این فرض درست نباشد (۱۸). از آنجا که روش کامل سازی معمولاً بر اساس داده‌های ایستگاه‌های اطراف است، مشاهده شد زمانی که فاکتور پیشگو گمشده باشد، تولید داده بر اساس روش‌های رگرسیون‌گیری ممکن است غیر قابل انجام باشد (۴). گذشته از آن معمولاً روش‌های کلاسیک رگرسیون‌گیری، یک متغیر وابسته را آنالیز می‌کنند، لذا برای تعداد زیادی از متغیرها، تولید معادلات رگرسیون‌گیری متفاوت می‌تواند بسیار وقت‌گیر باشد.

محدودیت‌های موجود در روش‌های سنتی پیدا کردن داده‌های گم شده که در بالا به آنها اشاره شد، موجب شد تا از شبکه‌های عصبی مصنوعی (Artificial Neural Network) به‌صورت گسترده‌ای استفاده شود. با کمک این روش‌ها برای پیدا کردن داده‌های گمشده می‌توان به‌راحتی از الگوهای غیر خطی مدل‌های پیچیده استفاده کرد. نقدی (۲۶) با بازسازی داده‌های دبی حوضه آبریز کارون بزرگ به روش شبکه عصبی مصنوعی و مقایسه آن با روش‌های دیگر، گروه‌های بازسازی مختلف را برای این روش‌ها اولویت بندی کرد و درنهایت رگرسیون ساده را به‌عنوان بهترین روش بازسازی داده‌های دبی در کل حوضه کارون بزرگ معرفی کرد. دستورانی (۸) در تحقیقی کارایی هوش مصنوعی را در تخمین داده‌های مفقود شده هیدرولوژی ارزیابی کرد. به این منظور، نتایج حاصل از شبکه عصبی مصنوعی با نتایج روش نسبت نرمال و روش همبستگی بین ایستگاه‌ها را مورد بررسی و مقایسه قرار داد و با استفاده از شاخص‌های مقایسه کارایی مدل، نتایج ارائه شده توسط هر یک از روش‌ها را با مقادیر مشاهداتی مقایسه کرد. متین‌زاده و همکاران (۲۴) کارایی الگوریتم ترکیبی ژنتیک عصبی مصنوعی در بازسازی داده‌های حداکثر بارش ۲۴ ساعته در استان چهار محال و بختیاری را مورد ارزیابی قرار دادند. نتایج تحقیق آنها در کلیه نواحی آب و هوایی این استان، برتری روش ترکیبی ژنتیک- عصبی مصنوعی را نسبت به شبکه عصبی نشان داد. بن آیسیا و همکاران (۷) با بررسی جامع روش‌های مختلف تک متغیره و چندمتغیره در بازسازی داده‌های مفقود هیدرولوژیکی، با استفاده از روش جک- نایف نشان دادند که کاربرد روش‌های چندمتغیره، در مقایسه با روش‌های جایگزینی و روش‌های درون‌یابی، موجب بهبود عملکرد و بهبود نتایج می‌شود.

الگوریتم نگاشت خودسامانده (Self- Organizing Map) مدلی اکتشافی برای مجسم ساختن و کشف روابط خطی و غیرخطی در مجموعه داده‌ها با ابعاد زیاد، تحت الگوریتم یادگیری غیرنظارت شونده (Unsupervised ANNs) است (۱۹). تاکنون SOM در بسیاری از مطالعات هیدرولوژیکی به‌کار رفته است. عملکرد شبکه‌های عصبی مصنوعی نسبت به روش‌های سنتی به شکل معنی‌داری بهتر است (۹). به‌ویژه مطالعات اخیر نشان دادند که نگاشت‌های خودسامانده (SOM) که حالت بدون نظارت شبکه‌های عصبی مصنوعی است، در بخش منابع آب عملکرد بهتری نسبت به شبکه‌های عصبی مصنوعی چند لایه پرسپترون (Multilayer Perceptron ANN) از خود نشان می‌دهد (۱۶). همچنین نگاشت‌های خودسامانده در طی آموزش بسیار قدرتمندتر از شبکه‌های عصبی مصنوعی چند لایه پرسپترون ظاهر می‌شوند، چرا که شبکه‌های عصبی مصنوعی چند لایه پرسپترون برای آموزش به یک سری داده کامل نیاز دارند . لذا در صورت وجود داده‌های گمشده، لازم

است قبل از آموزش MLP-ANNs تخمین‌های اولیه در مورد این داده‌های گمشده صورت گیرد. اخیراً SOM مستقیماً به‌عنوان ابزار تخمین‌گر قوی و غیرحساس به داده‌های مفقود به‌کار گرفته شده است.

وال و همکاران (۲۵) به‌منظور تخمین داده‌های مفقود در سری‌های زمانی بارندگی و رواناب از روش SOM استفاده کردند. نتایج آنها نشان داد که این تکنیک از توانایی خوبی جهت برآورد داده‌های مفقود هیدرولوژیکی برخوردار است. راستوم و همکاران (۲۸) مدل شبکه عصبی بر اساس SOM به‌منظور پیش‌بینی سریع اکسیژن خواهی بیوشیمیایی پنج روزه (BOD_5) را ارائه کردند و مقادیر مفقود را با استفاده از نگاشت خودسامانده تخمین زدند. نتایج مدل ارائه شده توافق خوبی با مقادیر اندازه‌گیری شده توسط روش سنتی داشت.

در این مقاله، با به‌کارگیری تکنیک SOM برای توسعه و بررسی مدل تخمین‌گر داده‌های مفقود در سری‌های زمانی بارش و رواناب، دو سناریو با ورودی‌های متفاوت تعریف شد و کارایی دو سناریو ارائه شده با شاخص‌های ارزیابی، مقایسه شد.

مواد و روش‌ها

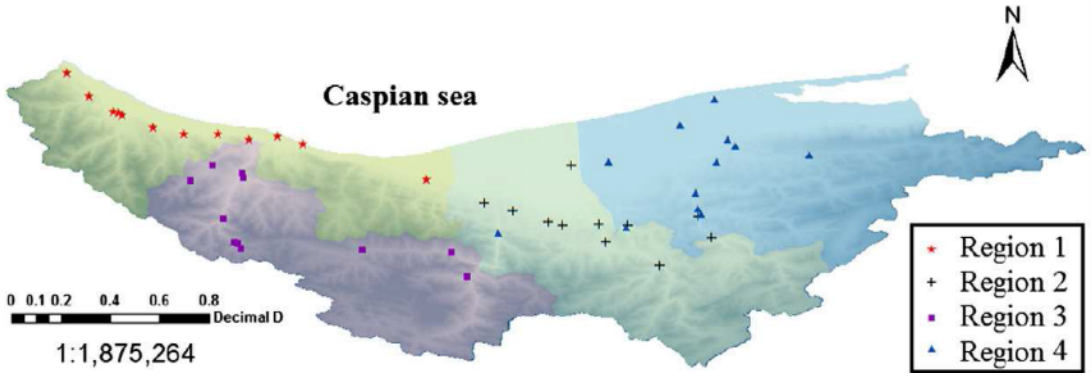
هدف از این تحقیق دستیابی به یک مدل هوش مصنوعی دقیق به‌منظور تخمین داده‌های مفقود، بدون توجه به نوع اقلیم و تنها با استفاده از سری‌های زمانی ناقص است. در ادامه منطقه مورد مطالعه و جزئیات الگوریتم به‌کار رفته تشریح می‌شود.

منطقه مورد مطالعه و داده‌های مورد استفاده

در این مطالعه به‌منظور بررسی کارایی روش نگاشت ویژگی خود سامانده در تخمین داده‌های مفقود در سری‌های زمانی بارندگی و دبی روزانه از ایستگاه‌های هیدرومتری و باران‌سنجی واقع در یک منطقه همگن از لحاظ هیدرولوژیکی در استان مازندران استفاده شد. کالته و همکاران (۱۷) توانایی SOM برای درون‌یابی داده‌های بارش را در یک منطقه با تغییرات مکانی و زمانی زیاد را در ایران مورد آزمون قرار دادند و به این نتیجه رسیدند که با آموزش روی داده‌های همگن، نتایج بهتری به‌دست می‌آید. لذا با توجه به آنکه استفاده از ایستگاه‌های هیدرومتری و باران‌سنجی واقع در یک منطقه همگن از لحاظ هیدرولوژیکی مزیت تلقی می‌شود، در این مطالعه از نتایج تحقیق فرسادنیا و همکاران (۱۱) به‌منظور انتخاب ایستگاه‌های ورودی الگوریتم SOM استفاده شد. فرساد نیا و همکاران (۱۱) کل حوضه‌های آبخیز استان مازندران را با استفاده از سه روش خوشه‌بندی سلسله مراتبی، فازی و شبکه‌عصبی مورد مطالعه قرار دادند و با استفاده از آزمون‌های سنجش همگنی هیدرولوژیکی بر پایه گشتاورهای خطی، استان مازندران را به چهار منطقه همگن از لحاظ هیدرولوژیکی تقسیم کردند (شکل ۱). در این مطالعه از اطلاعات ثبت شده در ۱۱ ایستگاه هیدرومتری و ۶ ایستگاه باران‌سنجی واقع در منطقه همگن شماره چهار استان مازندران (شکل ۱) استفاده شد. اسامی و مشخصات ایستگاه‌های مذکور به‌همراه درصد داده‌های گمشده مربوط به هر ایستگاه در جدول (۱) ارائه شده است. بیشتر داده‌های گمشده مربوط به داده‌های بارندگی است که در بازه ۱۱ درصد تا ۲۴ درصد قرار می‌گیرد. داده‌های مربوط به جریان کامل‌تر هستند و فقط ۵ ایستگاه دارای داده‌های گمشده هستند. البته الگوریتم پیش‌بینی SOM، تا زمانی که بردار ورودی مقدار داشته باشد (با وجود داده‌های گمشده) قادر به پیش‌بینی همزمان چندمتغیر است. تمام اطلاعات مورد نیاز از قبیل سری‌های زمانی بارندگی و جریان و همچنین موقعیت ایستگاه‌ها از شرکت مدیریت منابع آب ایران اخذ شد.

نگاشت خودسامانده (SOM)

نگاشت خودسامانده تابع چگالی احتمال از داده‌های ورودی تحت الگوریتم یادگیری غیرنظارت شونده است، که به‌عنوان روشی مؤثر برای خوشه‌بندی، خلاصه‌سازی و بصری کردن داده‌ها به‌کار می‌رود (۲۰). این الگوریتم دارای خصوصیات حفاظت از همسایگی و تجزیه و تحلیل فضای ورودی متناسب با توزیع داده‌ها است (۱۹ و ۲۰). SOM شامل دو لایه است:



شکل ۱. مناطق هیدرولوژیکی همگن استان مازندران (۱۱)

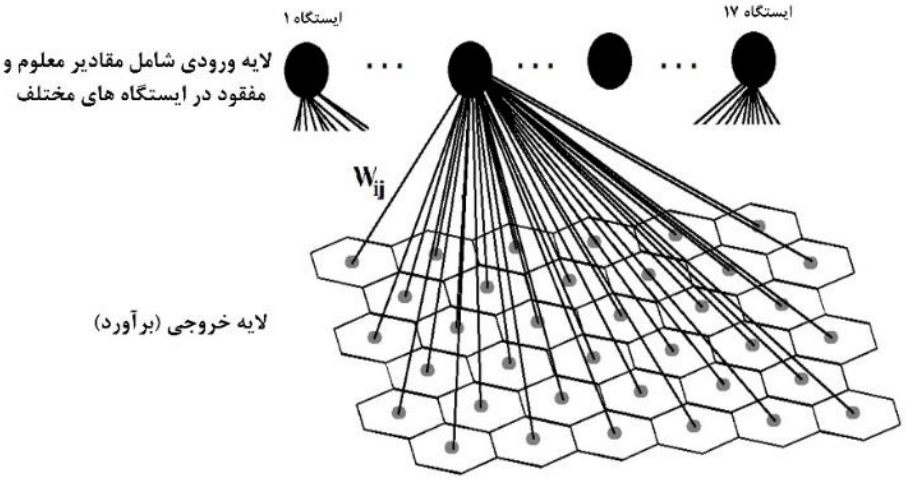
جدول ۱. مشخصات ایستگاه‌های مورد مطالعه

متغیر	کد ایستگاه	نام ایستگاه	رودخانه	حوضه آبریز	طول جغرافیایی (متر)	عرض جغرافیایی (متر)	ارتفاع ایستگاه (متر)	درصد داده‌های گمشده
بارش	۱۶-۰۸۵	والت	سردآبرود	سردآبرود	۵۱۹۰۳۱	۴۰۴۴۰۹۳	۹۷۵	۱۲/۵
	۱۶-۰۲۳	کلاردشت	سردآبرود	سردآبرود	۵۱۱۱۵۷	۴۰۳۷۴۶۵	۱۳۸۰	۰
	۱۵-۰۱۵	رزن	نور	هراز	۶۰۶۱۲۱	۴۰۰۶۳۸۶	۱۲۴۰	۰
	۱۵-۰۱۱	پنجاب	نمارستاق	هراز	۶۱۲۰۴۳	۳۹۹۵۳۲۶	۸۶۰	۴/۳
	۱۵-۰۱۳	بلده	نور	هراز	۵۷۳۷۶۶	۴۰۰۶۹۸۹	۱۳۶۰	۰
	۱۶-۰۱۷	دره هریجان	هریجان	چالوس	۵۲۸۳۶۲	۴۰۰۹۷۹۵	۱۹۰۰	۲۳/۴
	۱۶-۰۱۹	دوآب چالوس	هنیسک	چالوس	۵۳۰۲۱۲	۴۰۳۸۸۶۹	۳۷۵	۰
	۱۶-۰۲۱	پل ذغال	چالوس	چالوس	۵۲۹۸۸۰	۴۰۴۰۶۱۵	۳۵۰	۰
	۱۶-۰۷۹	پل مرگن	زنگوله	چالوس	۵۲۹۶۵۶	۴۰۰۷۰۹۷	۲۱۰۰	۳۳/۵
	۱۶-۰۸۱	ولی آباد	چالوس	چالوس	۵۲۷۱۹۶	۴۰۰۹۸۷۱	۱۷۵۰	۳/۸
	۱۶-۰۸۳	آبشار	چالوس	چالوس	۵۲۲۹۵۱	۴۰۲۰۶۳۴	۹۸۰	۰
بارش	۱۶-۰۸۵	والت	سردابرود	سردآبرود	۵۲۰۲۰۰	۴۰۴۲۴۵۴	۱۰۲۰	۱۴/۹
	۱۵-۰۰۹	نمارستاق	نمارستاق	هراز	۵۹۷۵۱۷	۳۹۹۵۵۸۳	۲۱۲۰	۱۸/۸
	۱۵-۰۱۱	پنجاب	نمارستاق	هراز	۶۱۴۳۶۲	۳۹۹۵۴۸۲	۹۲۰	۱۹/۲
	۱۵-۰۱۳	بلده	نور	هراز	۵۷۳۷۶۶	۴۰۰۶۹۸۹	۱۳۶۰	۲۳/۱
	۱۵-۰۱۵	رزن	نور	هراز	۶۰۶۱۲۱	۴۰۰۶۳۸۶	۱۲۴۰	۲۴/۵
	۱۶-۰۲۱	پل ذغال	چالوس	چالوس	۵۲۹۸۸۰	۴۰۴۰۶۱۵	۳۵۰	۱۰/۹

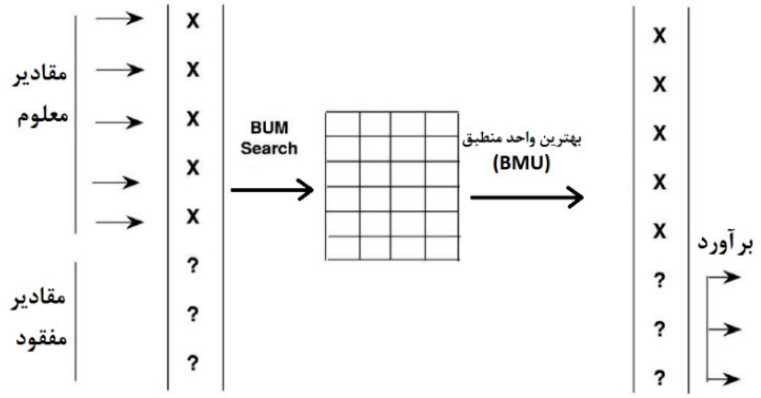
یک لایه ورودی متشکل از مجموعه گره‌ها (در این مطالعه داده‌های بارش و رواناب) و یک لایه خروجی (یا لایه کوهونن) که توسط گره‌هایی که در شبکه‌ای دو بعدی قرار گرفته‌اند، است (شکل ۲). تعداد بهینه واحدهای نقشه برابر $5 \times \sqrt{N}$ است که N تعداد بردارهای ورودی در مجموعه داده‌ها هستند.

با توجه به اینکه داده‌های خام ورودی شامل بارش و رواناب با واحدها و بزرگی‌های متفاوتی نسبت به یکدیگرند، از رابطه (۱) برای استاندارد سازی مقادیر ورودی به‌منظور بهبود مدل‌سازی استفاده می‌شود.

$$X' = (X_i - \overline{X}) / \sigma_x \qquad (۲)$$



شکل ۲. ساختار نگاشت ویژگی خود سامان (۱۲)



شکل ۳. پیش‌بینی مقادیر مفقود در بردار ورودی با استفاده از نگاشت خودسامانده (۲۷)

اگر $X_i(i=1,2,\ldots,m)$ یک متغیر ورودی باشد، $\bar{X}$ میانگین و σ_x انحراف معیار متغیر X_i هستند. استاندارد سازی داده‌ها باعث می‌شود تا میانگین مجموعه داده‌ها صفر و انحراف معیار آنها یک شود.

هر گره در لایه ورودی توسط شبکه سیناپسی به هر گره در لایه خروجی متصل است. هر گره خروجی دارای بردار ضرایب (وزن یا شدت اتصال) است که متصل به داده‌های ورودی است. زمانی که برداری از ورودی، به SOM وارد شود، گره‌ها در لایه خروجی با یکدیگر رقابت کرده و گره برنده (گره‌ای که فاصله تمامی وزن‌هایش از بردار ورودی حداقل است) انتخاب می‌شود. بر اساس قاعده یادگیری کوهونن، بردار وزن گره برنده و همسایه‌های از پیش تعریف شده‌اش در الگوریتم به‌روزرسانی می‌شوند. فرآیند یادگیری تا زمانی که معیار توقف (معمولاً زمانی که بردار وزن ثابت شود و/ یا زمانی که تعداد تکرارها کامل شود) معرفی شود، ادامه می‌یابد. در این روش هر گره در نگاشت توانایی تشخیص بردارهای ورودی مشابه خود را دارد. این ویژگی را خودسامان یابندگی می‌گویند، چرا که هیچ اطلاعات خارجی برای طبقه‌بندی نیاز نیست. برای جزئیات بیشتر در مورد الگوریتم SOM می‌توان به هایکین (۱۴) مراجعه کرد.

استفاده از SOM برای پیش‌بینی، که هدف اصلی این مطالعه است در شکل (۳) نشان داده شده است. ابتدا، مدل با استفاده از زیر مجموعه داده‌های موجود، آموزش داده می‌شود. سپس برداری ناقص یعنی بردار دارای مقادیر مفقود برای شناسایی بهترین واحد منطبق (Best Matching Unit)، بر اساس فاصله

اقلیدسی به مدل ارائه می‌شود. نرونی که بردار آن بیشترین مطابقت یا شباهت را با بردار داده‌های ورودی دارد به‌عنوان گره مورد نظر برگزیده خواهد شد که به آن بهترین واحد انطباق یافته (BMU) نیز گفته می‌شود. سپس مقادیر متغیر مفقود از مقدار متناظر آن در BMU، در نظر گرفته می‌شود.

ارزیابی کارایی مدل

کارایی مدل در مراحل آموزش و اعتبارسنجی با استفاده از معیارهای آماری زیر مورد بررسی قرار گرفت.

- ضریب تعیین (R^2 Coefficient of determination) بیانگر میزان احتمال همبستگی میان مقادیر مشاهده شده و پیش‌بینی شده است. مقدار R^2 بین صفر و یک متغیر است و مقدار R^2 برای پیش‌بینی مناسب باید نزدیک یک باشد.
- خطای مطلق میانگین (Mean absolute error) برای محاسبه متوسط خطا استفاده شد. همچنین ریشه میانگین مربعات خطا (Root Mean Square Error) تفاوت میان مقدار پیش‌بینی شده توسط مدل و مقدار واقعی است و برای مقایسه خطاهای پیش‌بینی توسط یک مجموعه داده است که در مقایسه با خطای مطلق میانگین وزن بیشتری به خطاهای بزرگ‌تر می‌دهد نیز برای اندازه‌گیری خطای کلی مدل استفاده شد.

$$MAE = \frac{\sum_{i=1}^{n} |P_i - O_i|}{n} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \quad (3)$$

در معادلات فوق P_i مقادیر پیش‌بینی شده و O_i مقادیر مشاهداتی است.

اطلاعات مورد استفاده به‌عنوان ورودی مدل

به‌منظور تهیه ورودی مدل تخمین‌گر داده‌های مفقود، ابتدا داده‌ها به‌صورت ستونی مرتب شدند، به‌نحوی که هر ستون دارای یک یا چند داده گمشده است. سپس دو سناریو تعریف شد، بدین‌ترتیب که در سناریو اول داده‌های ایستگاه‌های باران‌سنجی و دبی‌سنجی مورد بررسی به‌طور هم‌زمان و در سناریو دوم به‌صورت جداگانه به‌عنوان ورودی مدل استفاده شد.

نتایج و بحث

در این بخش به‌منظور مقایسه تأثیر ترکیب ورودی‌های مختلف در مدل SOM دو سناریو معرفی شد و نتایج آن با استفاده از آماره‌های ارزیابی، کارایی مدل مورد بررسی و تحلیل قرار گرفت.

اطلاعات مورد استفاده در سناریو اول

در این مطالعه ۱۷ متغیر (تعداد ایستگاه‌های اندازه‌گیری) وجود دارد. هر روز، یک بردار ورودی منفرد را تشکیل می‌دهد. بدین‌ترتیب ۹۴۹۷ بردار وجود دارد که مطابق با تعداد مشاهدات روزانه (بردارهای کامل و یا ناقص) است. بردارهای ورودی با داده مفقود یا بدون اطلاعات، در ستون‌ها با علامت NaN (یک عدد نیست یا Not a Number) مشخص شدند. مدل‌سازی شبکه عصبی غیرنظارت شونده خودسامانده توسط جعبه ابزار SOM در محیط نرم‌افزار متلب انجام گرفت. این جعبه ابزار توسط توسط آزمایشگاه اطلاعات و علوم کامپیوتر در دانشگاه هلسینکی گسترش یافته است. بر اساس ارتباطات چندمتغیره موجود بین دبی و بارش، تمام داده‌های متشکل از ۱۷ متغیر، به‌عنوان ورودی آموزش شبکه عصبی مورد استفاده قرار گرفت. این سناریو که تمام داده‌های دبی‌سنجی و باران‌سنجی به‌صورت یک‌جا به‌منظور تخمین داده‌های مفقود مورد استفاده قرار گرفتند، سناریوی اول نام‌گذاری شد.

اطلاعات مورد استفاده در سناریو دوم

در سناریو دوم داده‌های بارش و داده‌های دبی سنجی به‌صورت جداگانه به‌عنوان ورودی SOM مورد استفاده قرار گرفت. با استفاده از دو گروه داده مستقل برای آموزش شبکه، ۱۱ متغیر برای رواناب با ۹۴۹۷ بردار ورودی و ۶ متغیر برای باران سنجی با ۹۳۸۲ بردار ورودی، مورد استفاده قرار گرفت و عملکرد

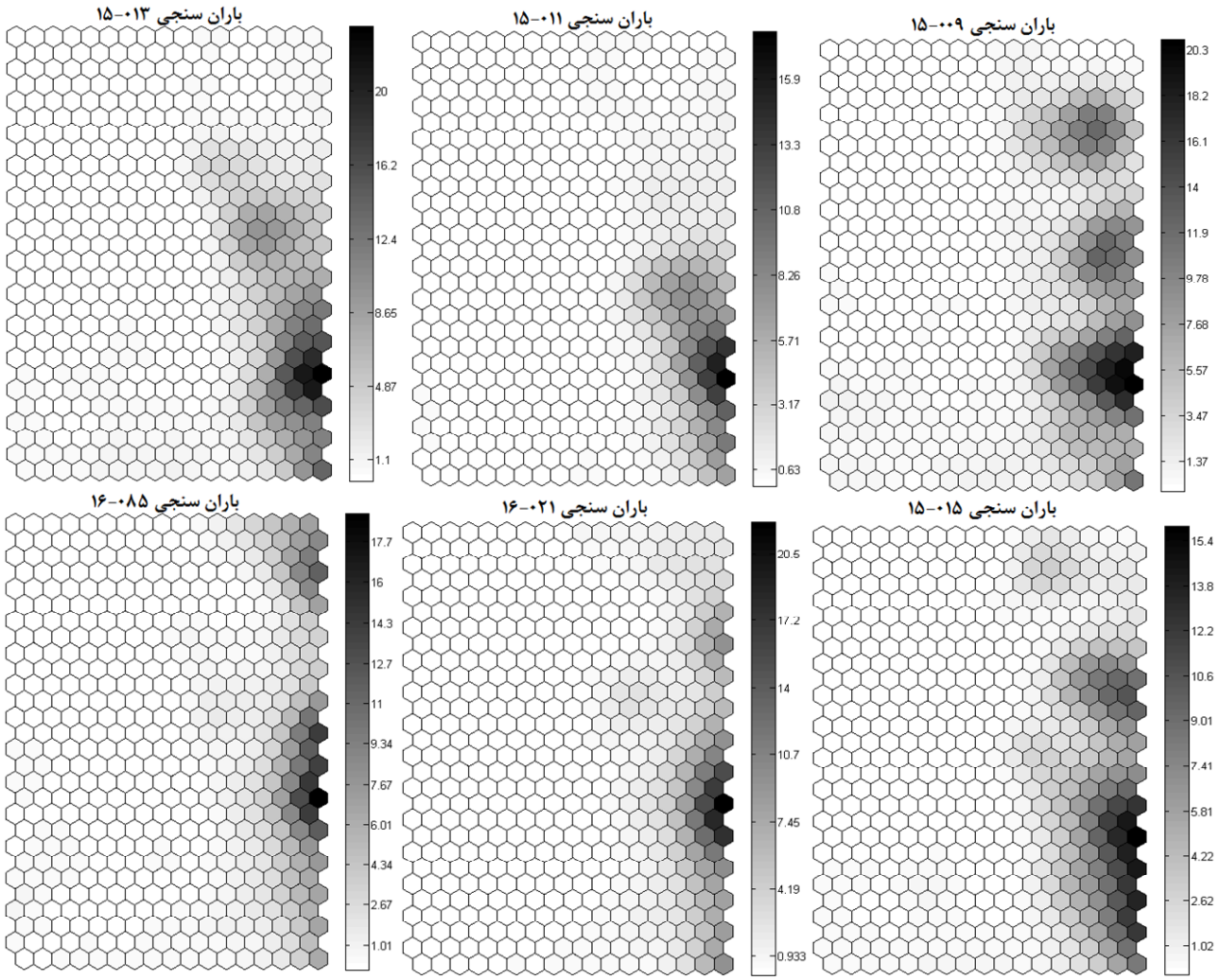
جدول ۲. معیارهای ارزیابی کارایی مدل در سناریوی اول

متغیر	نام ایستگاه	کد ایستگاه	ضریب همبستگی	ریشه میانگین مربعات خطا	خطای مطلق میانگین
ورودی	پل ذغال	۰۲۱– ۱۶	۰/۹۴	۴/۲۱	۲/۴۶
	کلاردشت	۰۲۳– ۱۶	۰/۹۳	۱/۲۲	۰/۷۱
	رزن	۰۱۵– ۱۵	۰/۹۲	۳/۳۲	۱/۶۹
	پنجاب	۰۱۱– ۱۵	۰/۷۳	۱/۵۵	۰/۶۳
	والت	۰۸۵– ۱۶	۰/۸۸	۲/۰۷	۰/۹۹
	بلده	۰۱۳– ۱۵	۰/۹۳	۲/۱۰	۱/۲۰
	دوآب چالوس	۰۱۹– ۱۶	۰/۸۷	۰/۸۱	۰/۴۵
	دره هریجان	۰۱۷– ۱۶	۰/۹۱	۰/۴۷	۰/۲۱
	پل مرگن	۰۷۹– ۱۶	۰/۸۹	۰/۵۳	۰/۲۲
	ولی آباد	۰۸۱– ۱۶	۰/۸۵	۱/۸۱	۰/۶۷
	آبشار	۰۸۳– ۱۶	۰/۹۴	۲/۶۱	۱/۴۴
خروجی	نمارستاق	۰۰۹– ۱۵	۰/۸۲	۲/۲۳	۰/۷۷
	پنجاب	۰۱۱– ۱۵	۰/۸۶	۱/۲۰	۰/۳۵
	بلده	۰۱۳– ۱۵	۰/۸۲	۱/۸۹	۰/۶۰
	رزن	۰۱۵– ۱۵	۰/۸۵	۱/۴۸	۰/۴۸
	پل ذغال	۰۲۱– ۱۶	۰/۸۹	۱/۴۶	۰/۵۳
	والت	۰۸۵– ۱۶	۰/۸۶	۱/۶۲	۰/۵۳

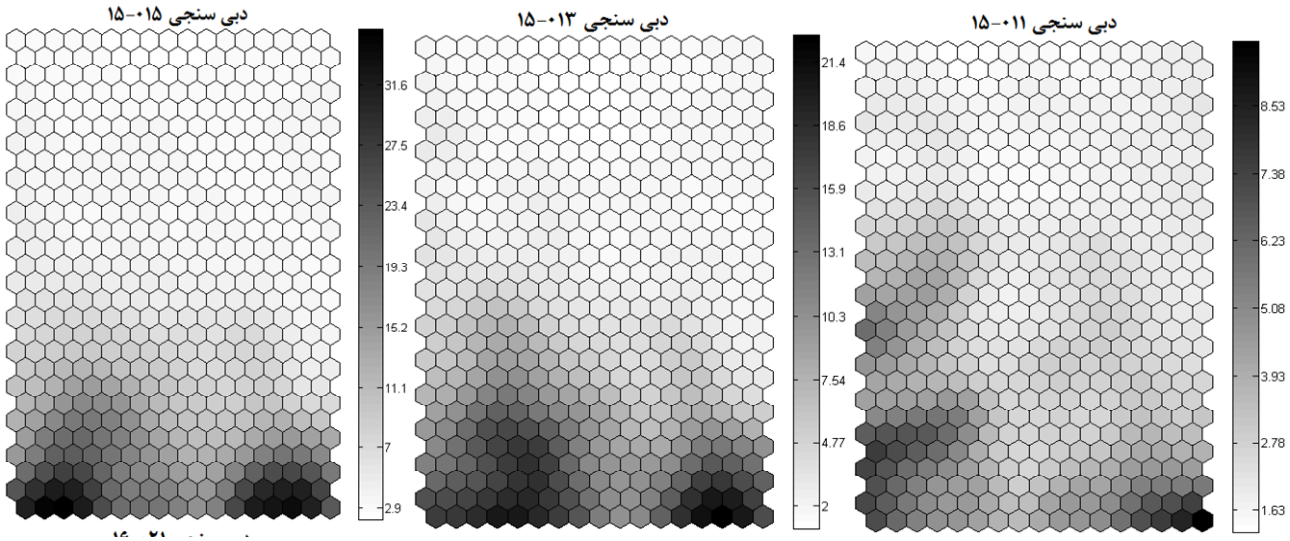
SOM از طریق ضریب همبستگی، خطای مطلق میانگین و ریشه میانگین مربعات خطا و همچنین بررسی بصری نمودارهای سری زمانی مورد ارزیابی قرار گرفت.

بررسی نتایج سناریوی اول

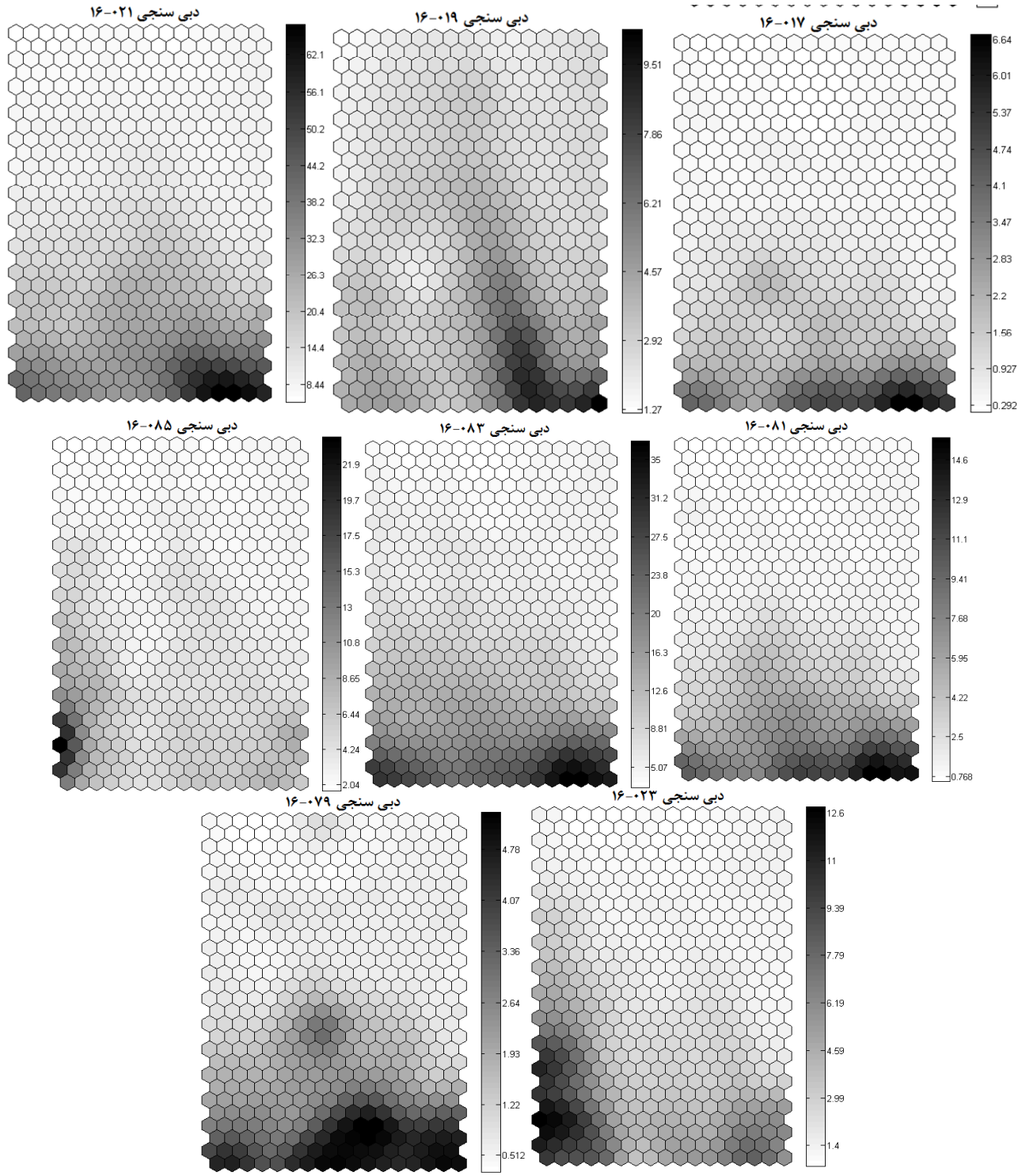
پس از آموزش شبکه SOM، تعداد بهینه واحدهای نقشه SOM معادل ۲۸×۱۷ به‌دست آمد. کیفیت آموزش SOM به‌وسیله خطای تدریجی میانگین کل (total average quantization error) و خطای توپوگرافیک کل (total topographic error) محاسبه شد. مقادیر نهایی خطای تدریجی میانگین کل و خطای توپوگرافیک کل به‌ترتیب برابر ۱/۱۹۳ و ۰/۰۸۸ به‌دست آمد. در بخش بعدی این مقادیر خطا با مقادیر خطای به‌دست آمده از گزینه دوم مقایسه می‌شود. عملکرد SOM با استفاده از ضرایب کارایی مدل برای سناریوی اول در جدول (۲) ارائه شده است. همان‌طور که مشاهده می‌شود (داده‌های بارش و جریان با یکدیگر آموزش دیده‌اند) توانایی SOM در پیش‌بینی بارش رضایت‌بخش نیست ولی در پیش‌بینی جریان خوب عمل کرده است. مقادیر ضریب همبستگی برای داده‌های جریان به‌جز ایستگاه ۰۱۱– ۱۵ که مقدار آن ۰/۷۳ است، در بیشتر ایستگاه‌ها حدود ۰/۸۵ یا بیشتر از آن است. همچنین مقادیر ضریب همبستگی برای ایستگاه‌های باران سنجی بین ۰/۸۲ تا ۰/۸۹ متغیر است. در سناریوی اول بین داده‌های جریان و بارش همبستگی وجود ندارد و یا مقدار آن ناچیز است. در شکل‌های (۴) و (۵) به‌ترتیب نمودارهای مؤلفه (Component) مربوط به داده‌های بارش و جریان در ایستگاه‌های مختلف نشان داده شده است. در این شکل‌ها به‌خوبی میزان همبستگی بین دو سری داده بارش و دبی مشخص هست. نمودار مؤلفه، نشان‌دهنده مقادیری از یک متغیر است که به‌وسیله هر واحد از نگاشت محاسبه شده است . لذا هر نمودار مؤلفه می‌تواند به‌عنوان بخشی از SOM فرض شود (۵). نمودارهای مؤلفه معمولاً به‌صورت رنگی یا خاکستری رنگ و شبکه‌های دوبعدی ترسیم می‌شوند. رنگ تیره نشان‌دهنده مقادیر زیاد هر مؤلفه و رنگ سفید کمترین مقدار را نشان می‌دهد. از این‌رو می‌توان از نمودارهای مؤلفه برای مقایسه بصری (از نظر همبستگی) بین



شکل ۴. نمودار مؤلفه‌های مربوط به داده‌های بارش در سناریوی (۱)



شکل ۵. نمودار مؤلفه‌های مربوط به داده‌های جریان در سناریوی (۱)



ادامه شکل ۵.

متغیرها استفاده کرد. برای مثال، اگر گرادیان یا شیب تغییر رنگ دو نمودار به یک صورت باشد (به‌صورت موازی تغییر کند) به معنای همبستگی مثبت و بالای دو متغیر نسبت به هم است، ولی اگر این شیب بر خلاف یکدیگر باشد به معنای همبستگی منفی این دو متغیر نسبت به هم است. همان‌طور که در شکل (۵) دیده می‌شود، همبستگی داده‌های جریان در ایستگاه‌های

جدول ۳. معیارهای ارزیابی کارایی مدل در سناریوی دوم

متغیر	نام ایستگاه	کد ایستگاه	ضریب همبستگی	ریشه میانگین مربعات خطا	خطای مطلق میانگین
C_{in}	پل ذغال	۱۶-۰۲۱	۰/۹۵	۳/۹۲	۲/۱۳
	کلاردشت	۱۶-۰۲۳	۰/۹۴	۱/۰۵	۰/۶۰
	رزن	۱۵-۰۱۵	۰/۹۴	۲/۷۹	۱/۴۳
	پنجاب	۱۵-۰۱۱	۰/۷۸	۱/۴۲	۰/۵۱
	والت	۱۶-۰۸۵	۰/۹۳	۱/۵۹	۰/۷۸
	بلده	۱۵-۰۱۳	۰/۹۴	۱/۹۴	۱/۰۴
	دوآب چالوس	۱۶-۰۱۹	۰/۹۰	۰/۷۵	۰/۳۷
	دره هریجان	۱۶-۰۱۷	۰/۹۲	۰/۴۴	۰/۱۹
	پل مرگن	۱۶-۰۷۹	۰/۹۲	۰/۴۵	۰/۱۸
	ولی آباد	۱۶-۰۸۱	۰/۸۵	۱/۷۷	۰/۶۰
	آبشار	۱۶-۰۸۳	۰/۹۴	۲/۴۶	۱/۲۶
C_{out}	نمارستاق	۱۵-۰۰۹	۰/۸۹	۱/۷۷	۰/۴۴
	پنجاب	۱۵-۰۱۱	۰/۸۹	۱/۰۸	۰/۲۵
	بلده	۱۵-۰۱۳	۰/۸۸	۱/۶۳	۰/۳۸
	رزن	۱۵-۰۱۵	۰/۹۰	۱/۲۳	۰/۳۰
	پل ذغال	۱۶-۰۲۱	۰/۹۲	۱/۲۲	۰/۳۳
	والت	۱۶-۰۸۵	۰/۹۱	۱/۲۹	۰/۳۲

مختلف، مشابه یکدیگر هستند. به‌عبارت دیگر، بیشترین مقادیر با رنگ تیره‌تر در سمت پایین نمودارهای مؤلفه (نمودارهای خروجی SOM) قرار گرفته‌اند. همچنین همبستگی داده‌های بارش در ایستگاه‌های باران سنجی مختلف نیز مشابه یکدیگر هستند، به این ترتیب که بیشترین مقادیر با رنگ تیره‌تر در سمت چپ نمودارهای مؤلفه (نمودارهای خروجی SOM) قرار گرفته‌اند. اما تفاوت بین نمودارهای مؤلفه جریان و نمودارهای مؤلفه بارش بسیار مشهود است. از طرفی این نتایج به همبستگی ضعیفی بین جریان و بارش نیز اشاره دارد. این همبستگی اگه چه بسیار ضعیف است ولی روی قابلیت پیش‌بینی SOM تأثیر گذاشته است. این موضوع قابلیت SOM در خوشه‌بندی و بصری کردن نتایج را نشان می‌دهد. به‌عبارت دیگر، الگوریتم SOM مقادیر مشابه را در نقشه خروجی SOM کنار هم قرار می‌دهد.

بعد از آموزش SOM و به‌دست آمدن وزن‌های نهایی، از بردار وزن‌های خروجی SOM برای تخمین داده‌های مفقود استفاده شد. به‌منظور بهبود بخشیدن به ضریب همبستگی در ایستگاه‌ها، داده‌های بارش و جریان جداگانه به SOM آموزش داده خواهند شد که تحت سناریوی (۲) در ادامه مورد بررسی قرار خواهد گرفت.

بررسی نتایج سناریوی دوم

در سناریوی دوم جهت بهبود نتایج، خصوصاً در مورد بارش، آموزش به‌صورت جداگانه انجام شد. آموزش اول روی داده‌های جریان و آموزش دوم روی داده‌های بارش که عملکرد SOM در جدول (۳) دیده می‌شود. پس از آموزش شکه SOM، تعداد بهینه واحدهای نقشه SOM برای ایستگاه‌های دبی‌سنجی برای ایستگاه‌های باران‌سنجی ۱۴×۳۵ به‌دست آمد. کیفیت آموزش SOM به‌وسیله خطای تدریجی میانگین کل و خطای توپوگرافیک کل سنجیده شد. مقادیر نهایی خطای تدریجی

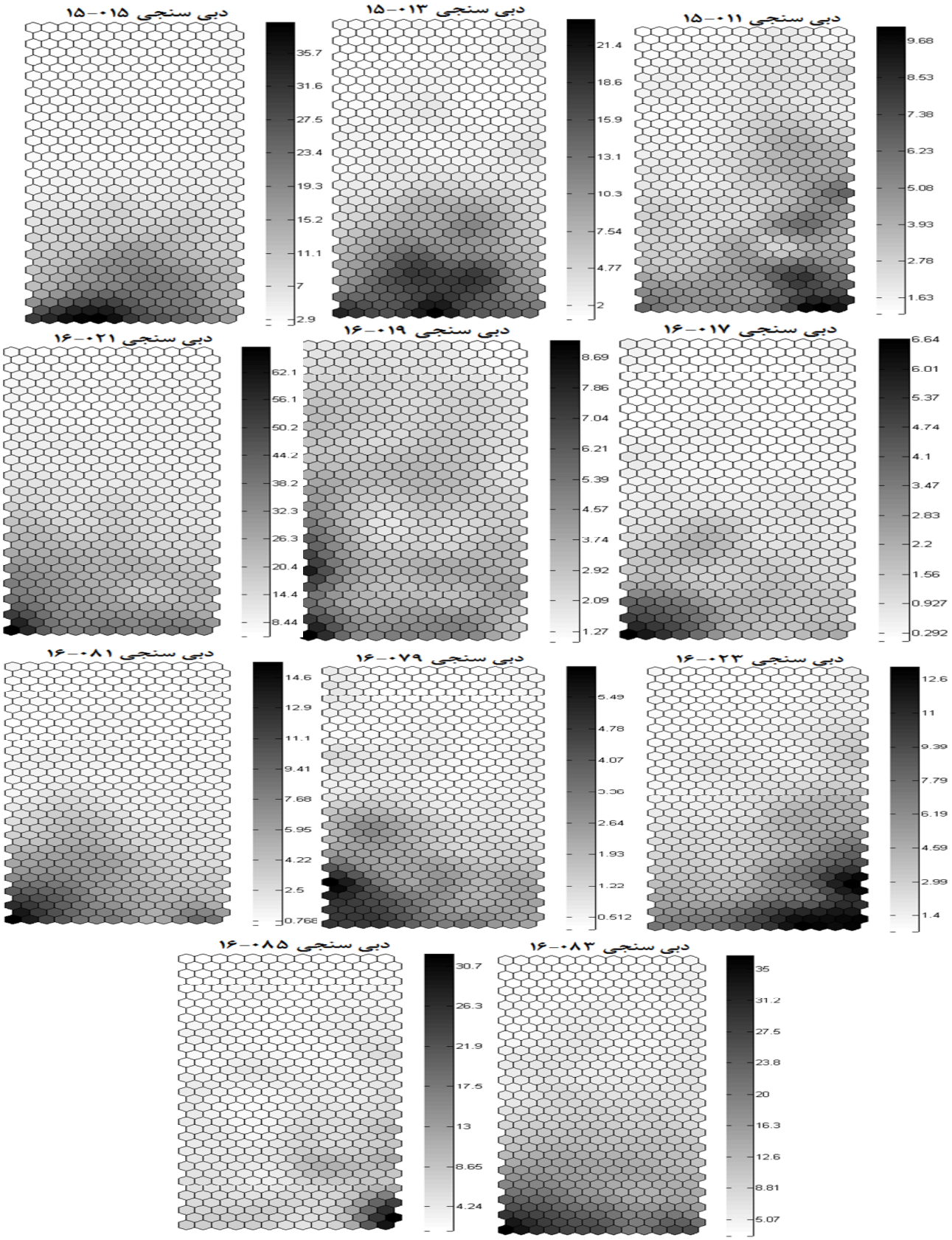
میانگین کل و خطای توپوگرافیک کل برای ایستگاه‌های دبی‌سنجی به‌ترتیب برابر ۰/۷۸۶ و ۰/۰۸۲ و برای ایستگاه‌های باران‌سنجی ۰/۳۰۹ و ۰/۶۰۲ به‌دست آمد. با مقایسه مقادیر خطاها با بخش (۳- ۳) مشخص است که میزان خطای آموزش مدل کاهش یافته است. به‌عبارت دیگر، دقت نتایج بهبود یافته است.

همان‌طور که در جدول (۳) مشاهده می‌شود، سازگاری بین داده‌های اولیه و داده‌های پیش‌بینی شده توسط SOM بهبود کلی یافته است. این موضوع هم برای داده‌های جریان و هم برای داده‌های بارش صادق است. عملکرد SOM در پیش‌بینی داده‌های جریان که با (R^2) مورد ارزیابی قرار گرفته است. برای مثال، برای ضریب همبستگی ایستگاه دبی‌سنجی پنجاب (۱۵- ۰۱۱) که در سناریو اول کمترین مقدار ضریب همبستگی را داشت، به مقدار ۰/۷۸ رسیده است. ایستگاه ۰۱۱- ۱۵ در این سناریو نیز کمترین مقدار را به‌خود اختصاص داد ولی این بار مقدار آن که برابر ۰/۷۸ است، که بهبود چشمگیری است. همچنین مقادیر ریشه میانگین مربعات خطا و خطای مطلق میانگین برای هر دو سناریو محاسبه و در جداول (۲) و (۳) ارائه شده است. همان‌طور که مشاهده می‌شود کارایی مدل در سناریو دوم بهبود چشمگیری داشته است و مقادیر RMSE و MAE برای تمامی ایستگاه‌ها کاهش یافته است. نمودارهای مؤلفه مربوط به داده‌های جریان در شکل (۶) ارائه شده است. همچون سناریو اول، در این سناریو نیز شباهت بین متغیرها وجود دارد، این شباهت بین بعضی از متغیرها مثل ایستگاه‌های ۰۷۹- ۱۶، ۰۱۷- ۱۶، ۰۲۱- ۱۶ و ۰۱۵- ۱۵ از نظر بصری آشکارتر است. برای سایر متغیرها پیدا کردن شباهت تا این مقدار از نظر بصری کار مشکلی است ولی با وجود این اختلافات، نتایج SOM در پیش‌بینی داده‌های گمشده جریان، رضایت‌بخش است. دامنه تغییرات مقادیر (R^2) برای داده‌های بارش در سناریوی دوم بین ۰/۸۹ تا ۰/۹۲ برآورد شد. نتیجه نمودارهای مؤلفه مربوط به بارش در شکل (۷) دیده می‌شود. در اینجا نیز شباهت‌های زیادی وجود دارد و می‌توان نتیجه گرفت که توانایی پیش‌بینی SOM در ارتباط با داده‌های گمشده بارش نیز رضایت‌بخش است. برای سناریوی (۲) که سناریو بهتر است، گرافی برای مقایسه بین داده‌های اصلی و پیش‌بینی شده توسط SOM رسم شد. این گراف برای داده‌های جریان، نسبت به ایستگاهی که بیشترین داده‌های گمشده را داشت رسم شد. این ایستگاه، ایستگاه هیدرومتری پل مرگن است که با کد ۰۷۹- ۱۶ شناخته شده و دارای ۳۳/۵ درصد داده گمشده است.

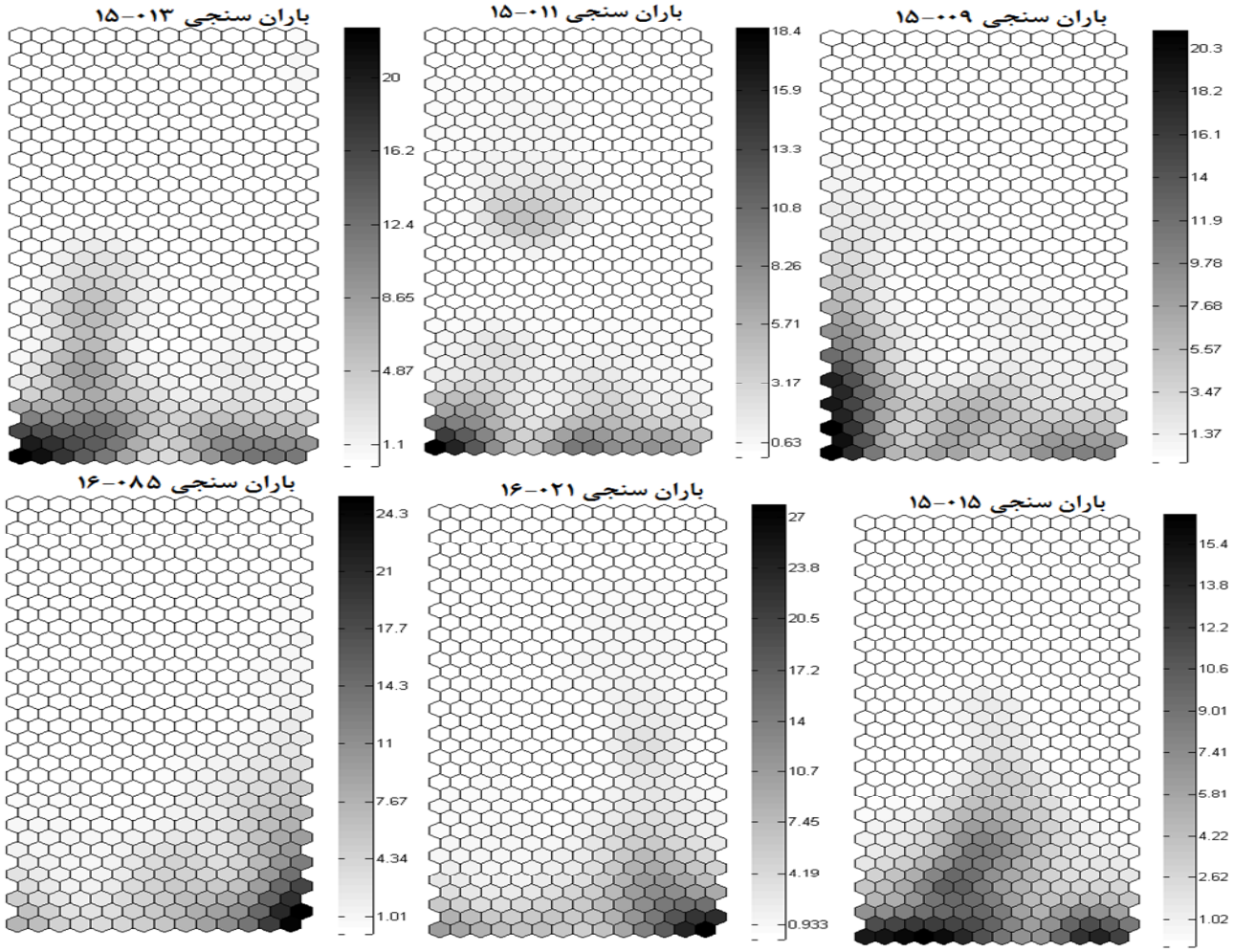
شکل (۸) گراف مربوط به مقایسه داده‌ها اصلی و پیش‌بینی شده جریان مربوط به ایستگاه پل مرگن با کد ۰۷۹- ۱۶ را نشان می‌دهد. برای رسم گراف بارندگی از داده‌های ایستگاه رزن با کد ۰۱۵- ۱۵ استفاده شد. این ایستگاه دارای ۲۴/۵ درصد داده گمشده است که در بین ایستگاه‌های باران‌سنجی بیشترین مقدار دیتای مفقود را دارد. شکل (۹) گراف مربوط به مقایسه داده‌های اصلی و پیش‌بینی شده بارش مربوط به ایستگاه رزن را نشان می‌دهد. بیشترین و کمترین مقادیر و همچنین روند تغییر داده‌های پیش‌بینی شده توسط SOM مطابقت خوبی با داده‌های اصلی دارد. نکته قابل تامل در نمودارهای (۷) و (۸) آن است که علیرغم داشتن ضریب همبستگی بالا میان داده‌های مشاهداتی و پیش‌بینی شده، الگوریتم SOM مقادیر حداکثری و نقاط پیک را کمتر از مقادیر مشاهداتی پیش‌بینی کرده است. لذا لازم است سری‌های زمانی تکمیل شده با استفاده از روش پیشنهادی SOM با احتیاط استفاده شود.

نتیجه‌گیری

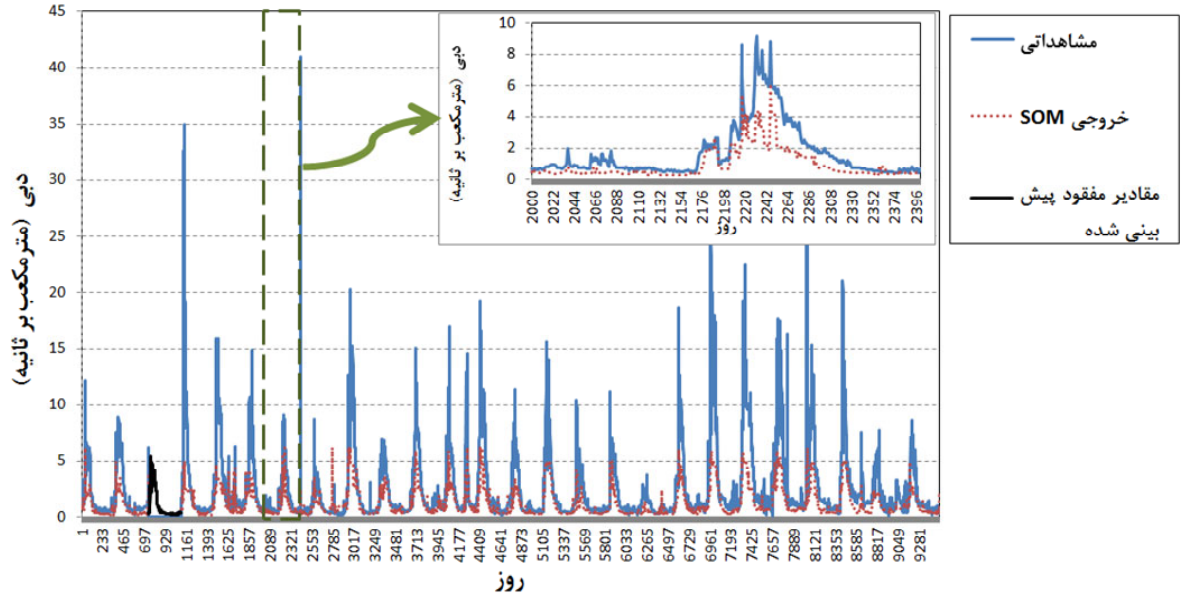
در حالت کلی SOM ابزار بسیار قدرتمندی برای پیش‌بینی تعداد زیادی متغیر با نسبت داده‌های گمشده بالا است. اگر چه کیفیت این پیش‌بینی به همبستگی داده‌های آموزشی بستگی دارد. همچنین این تحقیق نشان داد که در این منطقه قابلیت پیش‌بینی SOM در مورد داده‌های گمشده جریان، بهتر از داده‌های گمشده بارندگی عمل می‌کند. یک دلیل محتمل برای این رفتار، وجود تغییرات زیاد بین داده‌های بارش است که با نتایج مطالعه کالته و همکاران همخوانی دارد (۱۷). در واقع



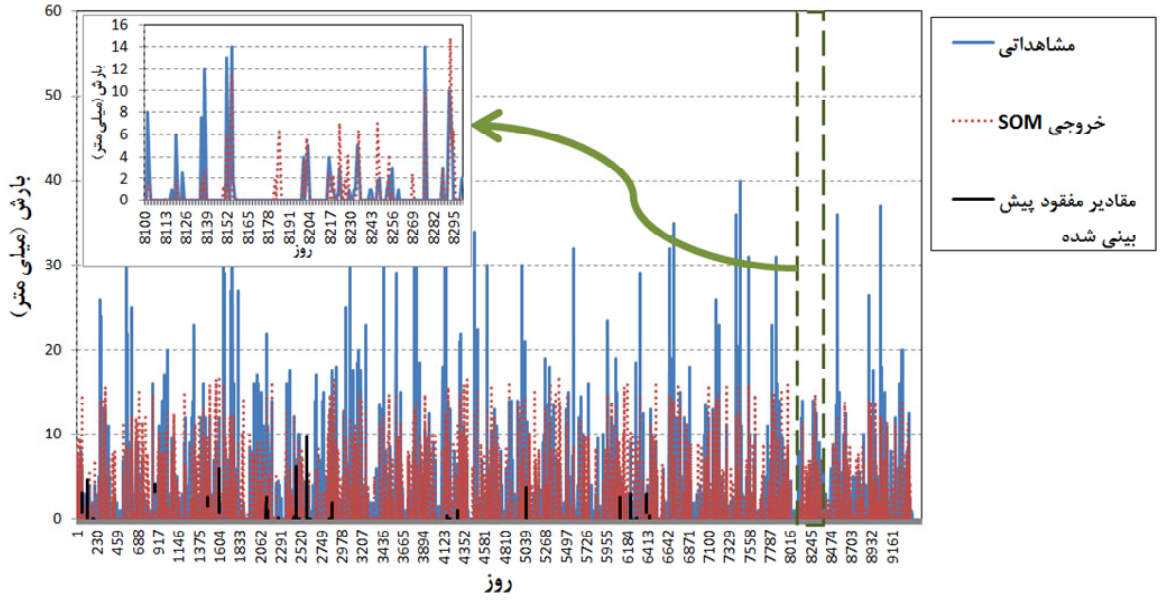
شکل ۶. نمودار مؤلفه‌های مربوط به داده‌های جریان در سناریوی (۲)



شکل ۷. نمودار مؤلفه‌های مربوط به داده‌های بارش در سناریوی (۲)



شکل ۸. مقایسه داده‌های اصلی و پیش‌بینی شده توسط SOM برای ایستگاه دبی‌سنجی پل مرگن



شکل ۹. مقایسه داده‌های اصلی و پیش‌بینی شده توسط SOM برای ایستگاه باران‌سنجی رزن

کارایی SOM زمانی که متغیرها تغییرات کمتری نسبت به یکدیگر داشته باشند و یا همبستگی بیشتری داشته باشند، بهبود می‌یابد.

نتایج این تحقیق نشان داد که الگوریتم SOM در پیش‌بینی نقاط پیک، کم برآورد (Under estimate) دارد. لذا لازم است در استفاده از سری‌های زمانی تکمیل شده با روش پیشنهادی SOM به کاربرد سری زمانی تکمیل شده توجه و در استفاده از سری زمانی تکمیل شده جانب احتیاط را لحاظ کرد.

پیش‌بینی بهتر جریان توسط SOM از دو جنبه بسیار حائز اهمیت است. یکی از مزایای این موضوع، ارجحیت داده‌های جریان برای ارزیابی مؤثر منابع آب است، همچنین اندازه‌گیری داده‌های جریان بسیار مشکل و پرهزینه است و با استفاده از این روش می‌توان سری‌های زمانی جریان را با دقت بالایی تکمیل کرد.

منابع مورد استفاده

1. Abdollahi, kh. 2006. Provide a new model and algorithm for the reconstruction of lost data. *In:* Proceeding of the Third National Conference on Erosion and Sediment, Iran.
2. Abatzoglou, J. T., K. T. Redmond and L. M. Edwards. 2009. Classification of regional climate variability in the state of California. *Journal of Applied Meteorology and Climatology* 48: 1527-1541.
3. Adeloye, A. J. 1996. An opportunity loss model for estimating value of stream flow data for reservoir planning. *Water Resources Management* 10(1): 45-79.
4. Adeloye, A. J. 2009. The relative utility of multiple regression and ANN models for rapidly predicting the capacity of water supply reservoirs. *Environmental Modelling & Software* 24(10): 1233-1240.
5. Adeloye, A. J. 2011. Reducing the uncertainty associated with water resources planning in a developing country basin with limited runoff data through AI rainfall–runoff modelling. *In:* Proceedings of the Symposium HS03 - Risk in Water Resources Management, Melbourne, Australia, IAHS 347, pp. 121-126.
6. Adeloye, A. J., R. Rustum. 2012. Self-organising map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins. *Hydrology Research* 43: 603-617.
7. Ben Aissia, M. A., F. Chebana and T. B. M. J. Ouarda. 2017. Multivariate missing data in hydrology – review and applications. *Advances in Water Resources* 110: 299-309. doi.org/10.1016/j.advwatres.2017.10.002.
8. Dastorani, M. T. 2007. Evaluation of the application of artificial intelligence model for simulation and real–time prediction of flood flow. *Journal of Water and Soil Scince* 11(40): 27-37.
9. Dastorani, M. T., A. Moghadamnia, J. Piri and M. Rico-Ramirez. 2010. Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment* 166: 421-434.

10. Dinpashoh, Y., D. Jhajharia, A. Fakheri-Fard, V. P. Singh and E. Kahya. 2011. Trends in reference crop evapotranspiration over Iran. *Journal of Hydrology* 399: 423-433.
11. Farsadnia, F., M. Rostami Kamrood, A. Moghaddam Nia, R. Modarres, M. T. Bray, D. Han and J. Sadatinejad. 2014. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. *Journal of Hydrology* 509: 387-397.
12. Fei, B. K. L., J. H. P. Eloff, M. S. Olivier and R. H. S. Venter. 2006. The use of self-organizing maps for anomalous behavior detection in a digital investigation. *Forensic Science International* 162: 33-37.
13. Gyau-Boake, P. and G. A. Schultz. 1994. Filling gaps in runoff time series in West Africa. *Hydrological Sciences Journal* 39(4): 621-636.
14. Haykin, S. 2003. Neural networks: A comprehensive foundation. Fourth Indian Reprint, Pearson Education, Singapore.
15. Ilunga, M. and D. Stephenson. 2005. Infilling stream flow data using feed-forward back propagation (BP) artificial neural networks: application of standard BP and Pseudo Mac Laurin power series BP techniques. *Water SA* 31(2): 171-176.
16. Kalteth, A. M. and P. Hjorth. 2009. Imputation of missing values in a precipitation–runoff process database. *Hydrology Research* 40(4): 420-432.
17. Kalteth, A. M. and R. Berndtsson. 2007. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). *Hydrological Sciences Journal* 52(2): 305-317.
18. Khalil, M., U. Panu and W. Lennox. 1998. Estimation of missing streamflows: A historical perspective. *In:* Proceeding of the Annual Conference of the Canadian Society for Civil Engineering. Halifax, Nova Scotia, pp 235-246.
19. Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43: 59-69.
20. Kohonen, T. 2001. Self-Organizing Maps. Springer, Berlin, Germany.
21. Kumambala, P. G. 2010. Sustainability of water resources development for Malawi with particular emphasis on north and central Malawi. PhD Thesis, University of Glassgow.
22. Lookzadeh, S. 2005.Evaluation of several different methods of reconstruction of statistical precipitation discharges in different time scales in central Alborz. MSc. Thesis, Tehran University.
23. Linacre, E. 1992. Climate data and resources - A Reference Guide. Routledge, London and New York.
24. Matinzadeh, M., R. Fattahi, M. Shayannejhad and KH. Abdollahi. 2013. Estimation and reconstruction of maximum 24-hour annual precipitation data using computational model of genetic algorithm and artificial neural networks. *Iran Watershed Management Science and Engineering* 7(22): 2013.
25. Mwale, F. D., A. J. Adeloye and R. Rustum. 2012 .Infilling of missing rainfall and stream flow data in the Shire River basin, Malawi - A self-organizing map approach. *Physics and Chemistry of the Earth* 50: 34-43.
26. Naghdi, R., M. Shayannejhad and S. J. Sadatinejhad. 2010. Reconstruction of the runoff data in the Karoon watershed basin by artificial neural networks and comparing it with other methods. *Journal of Watershed Management Research* 1(1): 59-73.
27. Rustum, R. and A. J. Adeloye. 2007. Replacing outliers and missing values from activated sludge data using Kohonen self-organizing map. *Journal of Environmental Engineering* 133(9): 909-916.
28. Rustum, R., A. J. Adeloye and M. Scholz. 2008. Applying Kohonen self-organizing map as a software sensor to predict biochemical oxygen demand. *Water Environment Research* 80(1): 32-40.
29. Sadatinejhad, S. J. 1998. Statistical comparison of precipitation data reconstruction methods in Isfahan province, Master's Degree Thesis, Tarbiat Modarres University.
30. Sadatinejhad, S. J., R. Naghdi and M. Shayannejhad. 2011. Application of Fuzzzy Linear Regression for predicting annual discharge missing data in hydrometric station compared with other conventional methods. *Journal of Water and Soil Conservation* 17(4): 67-86.

**Journal of Water and Soil Science (Science and Technology of Agriculture and Natural Resources)
Vol. 23, Special Issue of Flood and Soil Erosion, Winter 2020, Isfahan University of
Technology, Isfahan, Iran.**

Estimation of Missing Daily Precipitation and Runoff Using Self-Organizing Map (A Case Study: Mazandaran Province)

S. Eslami Jamal Abad[1], A. Sharafati[1*], E. Mohammadi Golafshani[1] and F. Farsadania[2]

(Received: December 24-2017 ; Accepted: May 27-2018)

Abstract

Expert aquatic designers face many problems; among these, in hydrology, defective occurrences in time-series can cause errors in the ultimate results of the study. This more often happens in the regions where the number of hydrometric and rain gauge stations is limited. In addition, assessing, developing and maintaining the use of water resources require accessible long-term and high-quality quality hydrological time-series. Thus, this necessitates correcting the statistical flaws and magnifies the importance of how to deal with the problems in the hydrological analyses. Statistical methods are, currently, used to infill data and statistical gaps. In this study, in order to introduce a multivariate method for estimating the missing data on rainfall and runoff, in a hydrologic homogeneous region in the Mazandaran province, self-organizing map methods were examined under two scenarios and some reliable estimates were obtained. In this regard, the correlation coefficients between the observational data and the model output were calculated for the precipitation data up to 0.92 and up to 0.95 for the runoff data. Therefore, to avoid the reduction of uncertainty caused by the inadequate data in water resource management, this method could be used.

Keywords: Unsupervised neural network, Infilling time series, Missing data

1. Civil Engineering Department, Science and Research Branch, Islamic Azad University, Tehran, Iran.
2. Department of Water Engineering, College of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran.
*: Corresponding Author, Email: asharafati@srbiau.ac.ir