УДК 551.465:551.509.5 DOI: 10.29006/1564-2291.JOR-2024.52(4).10

ПОТОЧЕЧНЫЕ И КОМПЛЕКСНЫЕ МЕРЫ КАЧЕСТВА В ИССЛЕДОВАНИЯХ АТМОСФЕРЫ И ОКЕАНА: ОБЗОР МЕТОДОВ И ПОДХОДОВ

В. Ю. Резвов[1,2], М. А. Криницкий[1,2], Н. Д. Тилинина[1]

[1] Институт океанологии им. П. П. Ширшова РАН
Россия, 117997, Москва, Нахимовский проспект, д. 36;
[2] Московский физико-технический институт
(национальный исследовательский университет),
Россия, 141701, Долгопрудный, Институтский пер., д. 9,
e-mail: rezvov.vyu@phystech.su

В науках об океане и атмосфере для описания качества результатов моделирования различного рода, включая численный прогноз погоды, статистическую коррекцию, повышение пространственного разрешения данных, используются различные обобщающие количественные показатели, называемые метриками, или мерами качества. Метрики дают представление о точности воспроизведения процессов моделями и позволяют сравнивать модели путем оценки неопределенности их результатов. В настоящей статье представлена наиболее общая классификация встречающихся в научной литературе метрик качества. Для каждой группы мер качества приведены примеры их использования в научных задачах. Помимо оценки традиционных поточечных метрик исследуются комплексные меры, рассматривающие различные аспекты сеточных данных. Среди таких специфических метрик выделяются меры с акцентом на пространственной структуре, внутренних корреляциях и неоднородностях прогнозируемых полей переменных, вероятностные методы проверки ансамблевых прогнозов. Отдельное внимание в данной работе также посвящено описанию феноменологических метрик и мер, основанных на редких и экстремальных явлениях.

Ключевые слова: метрики качества, сеточные данные, масштабирование, верификация, поточечные метрики, комплексные метрики, ансамблевый прогноз, феноменологические метрики

Введение

Изучение океана, погоды и климата характеризуется разрывом между теоретическим и прикладным аспектами численных моделей (Reed et al., 2022). Эта дилемма противопоставляет способность моделей обобщать разноплановую геофизическую информацию и неопределенность их результатов при непосредственном применении для конкретных задач. Совмещение этих двух аспектов важно для повышения практической значимости исследований в науках об атмосфере и океане (Barsugli et al., 2013; Briley et al., 2020; Jagannathan et al., 2020; Moss et al., 2019; Reed et al., 2022).

Решению задачи объективной оценки качества моделирования может способствовать разработка так называемых метрик, или мер качества – количественных

показателей, упрощающих сравнение сложных многомерных наборов данных до кратких числовых или категориальных представлений в легко интерпретируемом виде (Jagannathan et al., 2020; Reed et al., 2022).

Метрики дают понимание того, насколько хорошо та или иная модель или набор данных в целом описывает ключевые геофизические явления и процессы, имеющие первостепенное значение для конкретной задачи. Метрики также обеспечивают сравнение между собой различных наборов данных (Reed et al., 2022).

Кроме того, правильная оценка необходима для количественного описания неопределенности в самих данных, используемых для построения моделей (Tian et al., 2016). Оценка неопределенности позволяет также разделить систематическую и случайную ошибки модели (Barnston, Thomas, 1983; Ebert, McBride, 2000; Tian et al., 2016).

Целью настоящей работы является краткая систематизация наиболее часто применяемых типов метрик качества в задачах атмосферы и океана с описанием их преимуществ и недостатков на основе имеющейся научной литературы.

Раздел 1 посвящен описанию общих свойств метрик применительно к сеточным геофизическим данным. Раздел 2 предлагает наиболее общую классификацию мер качества. В разделе 3 рассматриваются применяемые для непрерывных сеточных данных поточечные метрики, а также ограничения, присущие таким классическим метрикам. Раздел 4 описывает более специализированные комплексные метрики, включая метрики для бинарных и категориальных данных, меры, основанные на сравнении распределений вероятностей, метрики для ансамблевых прогнозов, меры, разработанные с учетом пространственной корреляции, и феноменологические меры, включая оценку воспроизводимости экстремальных событий.

1. Общие свойства метрик в исследованиях атмосферы и океана

Важность выбора метрик качества объясняется их повсеместным применением к сеточным геофизическим данным (Sun et al., 2024). Сеточные данные представляют собой значения физических переменных, например, осадков, приповерхностной температуры, скорости ветра, приповерхностной солености и др. в узлах пространственных сеток. Сеточные данные обычно получаются путем комбинации измерений метеостанций, данных дистанционного зондирования, атмосферных моделей и реанализов (Yeggina et al., 2020; Zandler et al., 2020; Song et al., 2022; Sun et al., 2024).

Одним из примеров проблем, возникающих при верификации сеточных геофизических данных, является необходимость сопоставления пространственного поля на сетке с наблюдениями, полученными в нерегулярно расположенных точках. Упрощение применения метрик качества может быть результатом, например, ассимиляции данных для оперативных моделей численного прогноза погоды, что обеспечивает наличие верифицирующего поля на той же сетке, с той же пространственной и временной разрешающей способностью, что и прогнозное поле (Jolliffe, Stephenson, 2012),

однако такой подход не универсален. Например, модели, прогнозирующие осадки, недостаточно хорошо усваивают данные, а сильная пространственная изменчивость осадков, особенно в районах со сложным рельефом, а также сильная неравномерность расположения и плотности наблюдательных станций, усложняют анализ данных об осадках (Jolliffe, Stephenson, 2012).

Несмотря на кажущуюся простоту определения понятия «метрика качества», не существует метрик, одновременно обобщающих и анализирующих все данные, содержащиеся в наборе пространственных полей переменных (Reed et al., 2022). Таким образом, применение только одной метрики часто оказывается недостаточным. Набор метрик, анализирующих множество таких аспектов, как пространственные распределения переменных или временные ряды, позволяет всесторонне изучать, например, результат моделирования (Reed et al., 2022).

Сложные наборы данных также могут быть подвергнуты методам сокращения размерности для создания обобщенной интерпретируемой метрики (Collier et al., 2018; Taylor, 2001; Wehner, 2013).

Правомерность применения тех или иных метрик для оценки качества модели в целом зависит от предположения о статистической стационарности ошибок измерений или ошибок модели. При вычислении набора метрик необходимо выбирать соответствующие пространственные и временные области, в пределах каждой из которых характеристики ошибки модели считаются стационарными, и обрабатывать каждую такую область (например, времена года) отдельно (Tian et al., 2016). Такое разделение полного набора данных на подмножества может быть выполнено несколькими способами (Jolliffe, Stephenson, 2012).

Важность сеточных данных в науках об океане и атмосфере определяется применением в ряде задач, в том числе, для анализа изменений климата (Gilabert et al., 2015). Несмотря на существование высокоразрешенных региональных и даже глобальных реанализов и прогнозов погоды, характерные значения пространственного разрешения большинства атмосферных информационных продуктов, включая реанализы, прогнозы погоды и моделирование океана и климата, составляют десятки и даже сотни километров из-за экономии вычислительных мощностей (Kikumoto et al., 2015; Maina et al., 2020). Однако высокое пространственное разрешение данных часто имеет критическое значение, например, при прогнозировании экстремальных осадков, характеризующихся высокой пространственной неоднородностью.

Проблема противопоставления вычислительной эффективности и требуемого пространственного разрешения результатов численного моделирования может быть решена с помощью специальных методов повышения их разрешения. В англоязычной литературе такие методы объединены под общим названием *downscaling* (Hewitson, Crane, 1996; Wilby, Wigley, 1997; Schoof, 2013; Sun et al., 2024). Преобразование данных с грубым разрешением в их представление с высокой детализацией повышает полезность сеточных данных для различных применений.

Поскольку русскоязычная терминология в данной прикладной области наук о Земле еще не устоялась, вместо калькированного «даунскейлинг» мы будем

пользоваться более привычным для русского языка термином «пространственное масштабирование», понимая под ним не изменение пространственного разрешения данных вообще, а именно повышение такого разрешения.

Мы полагаем, что исследование применимости метрик качества в настоящий момент чаще всего продвигается в контексте пространственного масштабирования полей геофизических переменных, хотя и не ограничивается этой задачей. Это объясняется тем, что высокочастотная пространственная изменчивость полей переменных, например, в прогнозах погоды с высоким разрешением, представляет собой сложную проблему для проверки качества прогноза, мотивируя на создание множества новых мер качества (Gilleland et al., 2009).

2. Классификация метрик

В настоящей работе мы предлагаем наиболее общую классификацию метрик качества, подразделяющую их на две категории (Sun et al., 2024).

В первую категорию входят показатели, основанные на поточечных ошибках между значениями, спрогнозированными моделью, и фактическими значениями переменных. Примерами таких метрик являются средняя арифметическая ошибка (*B*), средняя абсолютная ошибка (*MAE*), средний квадрат ошибки (*MSE*), средняя квадратическая ошибка (*RMSE*), коэффициент корреляции (*r*) и коэффициент детерминации (R^2). К этой же группе относится и пиковое отношение сигнала к шуму (*PSNR*), рассчитываемое по максимальному значению оцениваемой величины и среднему квадрату ошибки.

Чаще всего, такие поточечные метрики применяются к непрерывным величинам, например, атмосферному давлению, температуре, толщине льда. Непрерывные величины теоретически могут принимать любые значения в пределах заданного диапазона, т. е. в любом интервале существует бесконечное множество их возможных значений. Однако на практике такие переменные часто рассматриваются как дискретные, или принимающие счетное количество значений, поскольку точность измерительных приборов ограничена, и данные содержат конечное число знаков после запятой (Jolliffe, Stephenson, 2012).

Во второй группе метрик вместо непосредственной ошибки значений в узлах сетки проводится оценка различных аспектов модели. Примерами таких метрик являются индекс структурной схожести (*SSIM*), эффективность Клинга-Гупты (*KGE*), оценка относительных операционных характеристик (*ROCSS*), смещение 98-го перцентиля (*P98*), оценка непрерывной ранжированной вероятности (*CRPS*), оценка по пространственным долям (*FSS*), критерий Пирси-Обухова (*PSS*) и др.

Среди прочего, в число таких комплексных метрик входят меры, оценивающие прогнозирующую способность в отношении экстремальных значений, и метрики, применимые к категориальным величинам – дискретным переменным, которые могут принимать только одно из конечного числа заранее заданных значений. Категории

таких величин могут быть упорядочены, как балл облачности, и неупорядочены, как, например, тип облачности. Бинарные переменные, имеющие только два возможных значения, представляют собой наиболее простую форму категориальных переменных; они указывают на наличие или отсутствие явления, например, дождя, тумана или грозы (Jolliffe, Stephenson, 2012).

Очевидно, такая классификация не является единственной или объективно оптимальной. Мы предлагаем опереться именно на рассмотренное выше деление метрик качества, поскольку оно отражает сложившуюся в задачах с геофизическими данными ситуацию. Как будет упомянуто позднее, несколько классических мер из группы поточечных метрик используются в подавляющем большинстве работ, а многие более сложные комплексные метрики встречаются исчезающе редко.

Тем не менее, мы считаем важным упомянуть и более подробную классификацию комплексных мер. Например, по результатам исследований, проведенных в рамках проекта по сравнению методов верификации пространственных прогнозов (*Spatial Forecast Verification Methods Inter-Comparison Project, ICP*), комплексные метрики качества было предложено сгруппировать в четыре подгруппы (Gilleland et al., 2009): (1) окрестные метрики (*neighborhood, or fuzzy*), (2) метрики разложения по масштабу (*scale separation, or scale decomposition*), (3) метрики, основанные на особенностях, или феноменологические метрики (*features based, or object based*) и (4) метрики деформации поля (*field deformation*).

Первые две подгруппы объединяются в более широкую группу фильтрующих метрик, поскольку и окрестные метрики, и метрики разложения по масштабу применяют пространственный фильтр к одному или обоим полям геофизических переменных (или к их разности), а затем вычисляют статистику по отфильтрованным полям. Фильтр обычно позволяет определить характерные пространственные масштабы, в которых прогноз считается качественным. Окрестные метрики применяют сглаживающий фильтр, а метрики разложения по масштабу применяют несколько полосовых пространственных фильтров, например, Фурье, вейвлеты и т. д. (Gilleland et al., 2009).

Третью и четвертую подгруппы метрик качества можно объединить в группу метрик смещения, поскольку, в отличие от поиска масштаба, на котором достигается желаемый уровень качества прогноза, эти меры описывают, какое смещение в пространстве требуется, чтобы сопоставить поле прогноза с фактическим полем (Gilleland et al., 2009).

Метрики деформации поля представляют собой определение такого пространственного преобразования прогнозного поля, которое минимизирует поточечную разницу, например, среднюю квадратическую ошибку между фактическим полем и «деформированным» прогнозом. Количественно метрика оценивает поле векторов искажения (*distortion vectors*), описывающих оптимальное пространственное преобразование прогнозного поля (Gilleland et al., 2009).

Феноменологические метрики будут описаны позднее в разделе 4.

Таким образом, несмотря на существование более подробных классификаций комплексных метрик, мы не делим совокупность комплексных метрик на

подгруппы, а считаем целесообразным перечислить наиболее часто применяемые комплексные метрики, исходя из основных недостатков поточечных метрик, рассмотренных в разделе 3.

3. Поточечные метрики

В таблице 1 перечислены основные поточечные метрики качества (Tian et al., 2016; Sun et al., 2024).

Табл. 1 – Основные поточечные метрики качества*

Метрика качества	Определение	Диапазон принимаемых значений	Идеальное значение
Средняя арифметическая ошибка *B*	$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\right)$	(–∞; +∞)	0
Средняя абсолютная ошибка *MAE*	$\frac{1}{N}\sum_{i=1}^{N}\left\|y_i - x_i\right\|$	[0; +∞)	0
Средний квадрат ошибки *MSE*	$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\right)^2$	[0; +∞)	0
Средняя квадратическая ошибка *RMSE*	$\sqrt{MSE}$	[0; +∞)	0
Коэффициент корреляции *r*	$\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}}$	[–1; 1]	1
Коэффициент детерминации R^2	$R^2 = 1 - FVU$ или $R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - x_i\right)^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$	(–∞; 1]	1
Пиковое отношение сигнала к шуму *PSNR*	$10 \lg\left(\frac{\left[\max_i\left\{x_i\right\}\right]^2}{MSE}\right)$	(–∞; +∞)	+∞

* Фактические значения и вывод модели обозначены как x и y, соответственно. Используются обозначения x_i и y_i, где $i = 1, \ldots, N$, для значений в i-м узле пространственной сетки $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$, $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2$, FVU – доля объясненной дисперсии.

Среди поточечных метрик выделяется «большая тройка» – средняя арифметическая ошибка, средняя квадратическая ошибка и коэффициент корреляции (Fisher, 1958; Tian et al., 2016; Wilks, 2011). «Большая тройка» метрик широко применяется в различных задачах, связанных с изучением океана и атмосферы (Taylor, 2001).

На правомерности их использования для оценки качества геофизических исследований акцент часто не делается. Тем более интересными, на наш взгляд, являются работы, анализирующие корректность и недостатки классических метрик со статистической или физической точек зрения. Анализ применения основных метрик обнаруживает ряд их принципиальных ограничений (Tian et al., 2016).

Классические метрики взаимосвязаны между собой и не могут считаться независимыми показателями качества метода (Murphy, 1988; Barnston, 1992; Taylor, 2001; Gupta et al., 2009; Entekhabi et al., 2010), что может указывать на избыточность одновременного их применения. Поточечные метрики качества также характеризуются тем, что различные комбинации ошибок между прогнозируемым и фактическим значениями переменных могут приводить к одинаковым значениям метрик (Tian et al., 2016).

Средняя квадратическая ошибка повсеместно используется в качестве стандартного статистического показателя для оценки качества модели в исследованиях в области метеорологии, климата и океана (McKeen et al., 2005; Savage et al., 2013; Chai, Draxler, 2014). Средняя абсолютная ошибка также широко используется при оценке моделей, хотя и не существует единого мнения относительно наиболее подходящей метрики для геофизических задач (Chai, Draxler, 2014). Неоднозначность средней квадратической ошибки, придающей ошибкам с более высокими абсолютными значениями больший вес в сравнении со средней абсолютной ошибкой, придающей одинаковый вес всем ошибкам, подчеркивается в работах (Willmott, Matsuura, 2005; Willmott et al., 2009).

Другие работы не внесли ясность в вопрос превосходства какой-либо из двух основных поточечных метрик. Chai, Draxler (2014) показали, что среднеквадратическую ошибку более целесообразно использовать, если ошибки модели соответствуют нормальному распределению, что также подтверждено в обсуждавшейся работе (Tian et al., 2016). В противном случае использование средней квадратической ошибки может привести к вводящим в заблуждение или бессмысленным результатам (Barnston, Thomas, 1983; Habib et al., 2001).

Если известны как фактические данные x, так и модельный прогноз y, то их совместное распределение $p(y,x)$ содержит всю не зависящую от времени информацию, относящуюся к проверке прогноза модели (Murphy et al., 1989; Tian et al., 2016). Поскольку совместное распределение выражается через условное распределение $p(y|x)$ и заранее известное распределение фактических данных $p(x)$ как $p(y,x) = p(y|x)\,p(x)$, то оценка качества модели через совместное распределение $p(y,x)$ равносильна оценке условного распределения $p(y|x)$ (Nearing, Gupta, 2015). Таким образом, определение модели ошибок $y(x)$ может задать необходимую поточечную метрику качества (Carroll et al., 2006; Wilks, 2011; Tian et al., 2016).

Коэффициент детерминации, обычно обозначаемый как R^2, используется в анализе дисперсии и регрессионном анализе. Он измеряет долю дисперсии, объясняемой моделью, и вычисляется по формуле $R^2 = 1 - FVU$, где FVU – это доля необъясненной дисперсии. Следовательно, более высокий показатель R^2 свидетельствует о том, что модель лучше соответствует данным.

В особом случае модели на сеточных данных, коэффициент детерминации R^2 определяется следующим образом:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - x_i \right)^2}{\sum_{i=1}^{N} \left(x_i - \overline{x} \right)^2},$$

где $\overline{x}$ представляет собой среднее фактическое значение, x_i – фактическое значение в i-м узле сетки, а y_i – смоделированное значение в i-м узле сетки. Если между фактическими и смоделированными переменными отсутствует связь, то $\overline{x}$ будет служить наилучшей «моделью», а слагаемые вида $(x_i - \overline{x})^2$ отражают отклонения от такой «модели».

Chai, Draxler (2014) доказали, что средняя квадратическая ошибка удовлетворяет неравенству треугольника – необходимому условию метрики расстояния, опровергнув критику, представленную в (Willmott, Matsuura, 2005). Средняя квадратическая ошибка проще для расчета градиента и поэтому более предпочтительна, так как ее квадрат может служить функцией потерь при оптимизации статистических моделей (Chai, Draxler, 2014).

Более поздняя работа (Hodson, 2022) делает вывод, что сравнивать среднюю квадратическую и среднюю абсолютную ошибку бессмысленно, поскольку они соответствуют различным распределениям ошибок. Авторы (Hodson, 2022) дополняют предложение (Chai, Draxler, 2014) об использовании комбинации различных показателей, включая среднюю квадратическую и среднюю абсолютную ошибки, для оценки качества модели. Метрики в комбинации должны быть взвешены по их правдоподобию, чтобы не нарушить теоретические принципы вывода метрик из соответствующих распределений ошибки модели. К обоснованному результату также можно прийти путем преобразования данных для соответствия тому или иному распределению, например, нормальному, что ведет к единственной поточечной метрике.

По мнению ряда авторов, для нелинейной модели ошибок классические метрики принципиально неприменимы и дают ошибочные интерпретации качества (Barnston, Thomas, 1983; Habib et al., 2001).

4. Комплексные метрики

Для пространственных полей геофизических переменных характерна когерентная пространственная структура с некоторыми неоднородностями, например, как в распределении осадков. Стандартные поточечные методы оценки качества часто не

учитывают присущую таким полям внутреннюю пространственную корреляцию. Как следствие, результат поточечных метрик часто сложно интерпретируем в физическом смысле (Casati et al., 2008).

Работа (Schultz et al., 2021) приходит к более широкому выводу, что поточечные метрики принципиально не подходят для применения в геофизических задачах. Количественная оценка качества модели прогнозирования погоды представляет собой более сложную задачу и требует разработки множества показателей, фокусирующихся на различных аспектах.

Таким образом, в ряде научных работ предлагается отойти от классических метрик, например, в сторону непараметрических метрик, особенно если невозможно определить параметрическую модель ошибок (Weijs et al., 2010; Gong et al., 2013; Vrugt, Sadegh, 2013; Nearing, Gupta, 2015), и в сторону более комплексных разносторонних метрик (Casati et al., 2008; Wilks, 2011), включая феноменологические показатели (Ebert, McBride, 2000).

В работах (Ebert, 2008; Weniger et al., 2017; Buschow et al., 2019) предлагаются различные методы оценки внутренней пространственной и временной корреляции в атмосферных данных. Другие метрики качества могут учитывать стохастическую природу специфических бинарных метеорологических переменных, например, наличие/отсутствие осадков (Roberts, Lean, 2008). Метрики, связанные с оценкой воспроизведения экстремальных событий, могут подвергнуть всю модель «дилемме прогнозиста», занижая оценку качества модели, выбирая только редкие по времени и пространству события (Lerch et al., 2017; Schultz et al., 2021).

Наиболее часто используемые в исследованиях атмосферы и океана комплексные метрики представлены в таблице 2.

Табл. 2 – Наиболее часто применяемые комплексные метрики

Метрика качества	**Диапазон принимаемых значений**	**Идеальное значение**	**Оцениваемые величины**
Индекс структурной схожести *SSIM*	[–1; 1]	1	Непрерывные величины
Эффективность Клинга-Гупты *KGE*	(–∞;1]	1	Непрерывные величины
Смещение 98-го перцентиля *P98*	(–∞; +∞)	0	Экстремальные значения непрерывных величин
Оценка непрерывной ранжированной вероятности *CRPS*	[0; +∞)	0	Вероятностные или ансамблевые прогнозы непрерывных величин
Оценка по пространственным долям *FSS*	[0; 1]	1	Бинарные величины
Оценка относительных операционных характеристик *ROCSS*	[0; 1]	1	Вероятностные прогнозы
Критерий Пирси-Обухова *PSS*	[–1; 1]	1	Бинарные величины
Критерий Багрова-Хайдке *PSS*	[–1; 1]	1	Бинарные величины

Индекс структурной схожести

Индекс структурной схожести позволяет аппроксимировать визуально воспринимаемую разницу в пространственной структуре между двумя изображениями (Wang et al., 2004). Эта метрика принимает значения от –1 до 1, при этом единица соответствует точному совпадению между двумя изображениями. Созданная для сравнения естественных изображений, метрика SSIM с помощью скользящего окна оценивает различия в их яркости, контрасте и структуре.

В обозначениях, применяемых в таблице 1, индекс структурной схожести рассчитывается для каждого положения скользящего окна:

$$SSIM = \frac{\left(2\overline{xy} + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\overline{x}^2 + \overline{y}^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)},$$

где σ_{xy} – ковариация x и y, а C_1 и C_2 – константы для предотвращения деления на ноль.

Повсеместность использования этой метрики качества и ее производных в задачах с атмосферными данными, в частности, для пространственного масштабирования метеорологических переменных (например, Ebtehaj et al., 2012; Geiss et al., 2022; Cho et al., 2024; Zhong et al., 2024), объясняется тем, что сеточные данные представимы в виде изображений, что превращает задачу масштабирования в задачу улучшения разрешения естественных изображений.

По сравнению с фактическим распределением метеорологической переменной, множество различных модельных реализаций полей могут иметь сходные показатели поточечных метрик, таких как средняя квадратическая ошибка, а их пространственные когерентные структуры – сильно отличаться (Ebtehaj et al., 2012). Индекс структурной схожести может учитывать ковариационную структуру ошибок в целом, что гипотетически устраняет некоторые общие недостатки поточечных метрик качества.

Тем не менее, в научной литературе растет обеспокоенность недостаточной обоснованностью применения индекса структурной схожести к сеточным геофизическим данным, в том числе из-за произвольности выбора масштаба скользящего окна (Robertson et al., 2014). Как следствие, некоторые авторы не рекомендуют использовать индекс структурной схожести в качестве полной замены поточечных метрик, так как этот индекс напрямую связан со средним квадратом ошибки, а его определение через средние значения, дисперсии и корреляции недостаточно хорошо моделирует восприятие зрительной системой человека (Dosselmann, Yang, 2011).

Эффективность Клинга-Гупты

Одним из возможных способов применения поточечных метрик в задачах с сеточными данными является комбинация таких метрик (Chai, Draxler, 2014). Исходя из таких же рассуждений, авторы (Gupta et al., 2009) предложили еще одну метрику качества модели – эффективность Клинга-Гупты, получившую распространение в

гидрометеорологических задачах, в том числе в задачах масштабирования (например, Shahriar et al., 2021; Böker et al., 2023).

KGE вычисляется следующим образом:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2},$$

где r – коэффициент корреляции, α – параметр изменчивости, β – параметр смещения.

В первоначальной форме, предложенной в (Gupta et al., 2009), параметр изменчивости выражается как отношение стандартных отклонений смоделированных и фактических данных $\alpha = \frac{\sigma_y}{\sigma_x}$, а параметр смещения – как отношение средних смоделированных и фактических значений $\beta = \frac{\overline{y}}{\overline{x}}$.

KGE может принимать значения от $-\infty$ до 1, при этом значение $KGE = 1$ соответствует идеальному согласию между моделированием и фактическими значениями.

Метрика *KGE* была модифицирована в работе (Kling et al., 2012). Поскольку параметры смещения и изменчивости могут быть взаимно скоррелированы, то вместо α было предложено использовать $\alpha_{\text{mod}} = \frac{\sigma_y / \overline{y}}{\sigma_x / \overline{x}} = \frac{\alpha}{\beta}$.

Несмотря на существование других модификаций эффективности Клинга-Гупты, например, непараметрической *KGE* (Pool et al., 2018), для всех вариантов этой метрики характерно общее свойство: несоответствие нуля «пограничному» значению, что некоторыми авторами рассматривается как препятствие. Другими словами, отрицательные значения не указывают на заведомо «плохое» качество модели, а положительные – на заведомо «хорошее», и поэтому значения *KGE* имеют лишь относительный смысл (Knoben et al., 2019).

Смещение 98-го перцентиля

Под 98-м перцентилем понимается значение, ниже которого расположены 98 % всех данных. При этом объединять данные можно различными способами: по всему временному интервалу для каждой отдельной точки пространственной сетки, по области исследования в целом и т. д.

В (Sun et al., 2024) утверждается, что само значение 98-го перцентиля можно считать метрикой качества модели. Действительно, *P98* является важной характеристикой для понимания и интерпретации распределения некоторой метеорологической переменной, особенно, «хвоста» с экстремально высокими значениями. Но на наш взгляд, правильнее говорить о смещении 98-го перцентиля, то есть о разности значений 98-го перцентиля одной и той же переменной между смоделированными и фактическими данными. Такая метрика применяется, в том числе, в задачах пространственного масштабирования атмосферных данных для определения того, насколько сильно смещено распределение высоких значений переменной в разных частях области моделирования относительно распределения фактических данных. Например, в (Hertig et al., 2019; Baño-Medina et al., 2021; Vaughan et al., 2022) смещение 98-го перцентиля

оценивается для моделей статистического масштабирования температуры и осадков, а в (O'Neill et al., 2017) – для масштабирования полей ветра.

Оценка непрерывной ранжированной вероятности

Оценка непрерывной ранжированной вероятности (*CRPS*) является одной из наиболее широко используемых метрик качества вероятностных или ансамблевых прогнозов геофизических переменных (Hersbach, 2000; Jin et al., 2023):

$$CRPS\left(\left\{y_i\right\}_{e=1,\ldots,E}, x_i\right) = \int_{s=-\infty}^{+\infty} \left(F_i(s) - \theta\left(s - x_i\right)\right)^2 ds,$$

где $\left\{y_i\right\}_{e=1,\ldots,E}$ – набор моделированных значений переменной в i-м узле сетки для каждого из E членов ансамбля, F_i – кумулятивная функция распределения соответствующей переменной в узле i, чаще всего, эмпирическая, полученная на основе ансамблевого прогноза, θ – функция Хевисайда, имеющая смысл кумулятивной функции распределения для прогноза, равного фактическому значению x_i (Jin et al., 2023).

Из определения *CRPS* следуют несколько ее полезных свойств. Во-первых, *CRPS* сосредоточена на всем допустимом диапазоне определенного параметра и рассматривается как ранжированная вероятностная оценка с бесконечным числом классов, каждый из которых имеет бесконечно малую ширину. Во-вторых, эта метрика чувствительна ко всему непрерывному диапазону интересующей переменной и не требует введения ряда предопределенных классов, от которых могут зависеть результаты. Кроме того, для детерминистического прогноза *CRPS* равна средней абсолютной ошибке и, следовательно, имеет четкую физическую интерпретацию (Hersbach, 2000).

CRPS учитывает как смещение, так и неопределенность прогноза членов ансамбля. Минимально возможное значение *CRPS* = 0 соответствует идеальному случаю, когда все прогнозы ансамбля идентичны наблюдению, а рост *CRPS*, неограниченный сверху, соответствует увеличению систематической ошибки и разброса ансамблевого прогноза (Jin et al., 2023).

Примеры использования оценки непрерывной ранжированной вероятности приведены также в (Bröcker, 2012; Alessi, DeGaetano, 2021; Harris et al., 2022).

Оценка по пространственным долям

В отличие от классических метрик, где оценивается соответствие прогноза и фактического значения в самой точке, существуют окрестные метрики – метрики, где оценивается прогноз и/или наблюдение в ее окрестности, что предполагает какую-либо обработку данных внутри скользящего окна. В качестве примеров такой обработки приводятся расчет среднего значения, наличие явления внутри скользящего окна, частота явления в окне, распределение величин в окне и др. (Бундель и др., 2021).

Одной из самых распространенных окрестных метрик качества является оценка по пространственным долям, разработанная изначально для оценки

моделирования осадков (Roberts, 2008; Roberts, Lean, 2008; Ebert et al., 2013; Skok, Roberts, 2018) без больших вычислительных затрат (Faggian et al., 2015). В оценке по пространственным долям рассматривается доля ячеек пространственной сетки в скользящем окне, занятых явлением, в поле прогноза P_y и в поле наблюдений P_x для всех положений скользящего окна. Далее рассчитывается средняя квадратическая разность долей по всем точкам сетки, деленная на худший возможный прогноз (Бундель и др., 2021):

$$FSS = 1 - \frac{\sum_{i=1}^{N}\left(P_{y,i} - P_{x,i}\right)^2}{\sum_{i=1}^{N} P_{x,i}^{\ 2} + \sum_{i=1}^{N} P_{y,i}^{\ 2}}.$$

Одним из наиболее важных свойств оценки по пространственным долям является способность определить пространственный масштаб, на котором модель можно считать имеющей приемлемое качество (Roberts, Lean, 2008; Ebert et al., 2013; Mittermaier et al., 2013; Skok, Roberts, 2018) за счет подбора размера скользящего окна. Кроме того, исследования показали, что метрика *FSS* связана с пространственным смещением осадков в прогнозе, что отчасти имитирует визуальную оценку качества прогноза осадков и позволяет легко интерпретировать значение *FSS* (Roberts, 2008; Skok, 2015).

Оцениваемая модель считается полезной, если $FSS > FSS_{uniform}$, где пороговое значение $FSS_{uniform} = 0.5 + f_0/2$, а f_0 – доля ячеек с явлением по наблюдениям во всей области прогноза. Идеальный прогноз имеет значение $FSS = 1$, а в наиболее плохом случае моделирования $FSS = 0$.

Хотя оценка по пространственным долям чаще применяется для оценки качества моделирования осадков (например, Glawion et al., 2023), она также может быть использована для других переменных, имеющих мелкомасштабные особенности, такие как приповерхностная температура воздуха (De Haan et al., 2015) и приповерхностный ветер (Gentile et al., 2022).

Характеристики оправдываемости прогноза

Еще одной важной категорией метрик качества, используемых в науке для оценки прогноза бинарных событий, являются характеристики оправдываемости прогноза. Для их расчета составляется таблица сопряженности. Пример приведен в таблице 3 для прогноза осадков, где n_{11} – число оправдавшихся прогнозов факта осадков; n_{12} – число неоправдавшихся прогнозов факта наличия осадков; n_{10} – число прогнозов наличия осадков; n_{21} – число неоправдавшихся прогнозов отсутствия осадков; n_{22} – число оправдавшихся прогнозов отсутствия осадков; n_{20} – число прогнозов отсутствия осадков; n_{01} – число случаев с осадками; n_{02} – число случаев без осадков; n_{00} – общее число прогнозов факта наличия и отсутствия осадков для данной выборки (Руководящий документ, 1991).

Табл. 3 – Таблица сопряженности для прогноза осадков

Прогноз	**Наблюдалось**		**Сумма**
	Осадки	**Без осадков**	
Осадки	n_{11}	n_{12}	n_{10}
Без осадков	n_{21}	n_{22}	n_{20}
Сумма	n_{01}	n_{02}	n_{00}

Согласно (Методические указания, 1991), качество прогноза осадков, а равно и других бинарных показателей, оценивается, в том числе, по критерию наличия или отсутствия оцениваемого явления. Это достигается путем сравнения прогноза в узлах регулярной сетки с данными, полученными на метеорологических станциях, находящихся не далее 50 км от узлов сетки или совпадающими с ними.

Для осадков при оценке прогнозирования рассчитываются следующие показатели (пример расчета для долей единицы):

общая оправдываемость прогноза осадков: $U = \dfrac{n_{11} + n_{22}}{n_{00}}$;

оправдываемость прогноза наличия осадков: $U_{\text{ос}} = \dfrac{n_{11}}{n_{10}}$;

предупрежденность факта наличия осадков: $\Pi_{\text{ос}} = \dfrac{n_{11}}{n_{01}}$;

оправдываемость прогноза факта отсутствия осадков: $\Pi_{\text{б-ос}} = \dfrac{n_{22}}{n_{02}}$;

предупрежденность факта отсутствия осадков: $U_{\text{б-ос}} = \dfrac{n_{22}}{n_{20}}$;

Критерий Багрова-Хайдке (*Heidke Skill Scor*e): $HSS = \dfrac{2\left(n_{11}n_{22} - n_{12}n_{21}\right)}{n_{10}n_{02} + n_{01}n_{20}}$.

Метрика HSS отражает долю оправдавшихся прогнозов после исключения из них тех, которые могли бы быть случайными. Значения критерия варьируются в диапазоне от –1 до 1. Положительные значения указывают на наличие полезного прогностического сигнала (Калмыкова, 2021).

Ключевым критерием качества прогноза является критерий Пирси-Обухова (*Peirce Skill Score*):

$$PSS = \frac{n_{11}}{n_{01}} - \frac{n_{12}}{n_{02}} \equiv \frac{n_{22}}{n_{02}} - \frac{n_{21}}{n_{01}}.$$

Данная метрика принимает значения от –1 до 1. Для идеального прогноза $PSS = 1$.

Как показано в (Ebert, Milne, 2022), из всех бинарных метрик только критерий Пирси-Обухова удовлетворяет всем трем критериям качества прогнозов погоды Мерфи, в особенности, для редких и экстремальных событий.

Критерии Мерфи определяют три различных типа качества прогнозов погоды. Прогнозы качественны, если (1) соответствуют суждению эксперта о качественном прогнозе; (2) соответствуют наблюдаемым погодным условиям в каждый момент

времени, на который сделан прогноз; и (3) предоставляют дополнительные экономические или иные выгоды лицам, принимающим решения на основе использования прогнозов (Murphy, 1993).

Феноменологические метрики

Как было сказано выше, главной проблемой классических метрик является то, что они не учитывают информацию о пространственной структуре полей геофизических переменных. Показательной, по нашему мнению, является ситуация, когда некая особенность или объект в прогнозируемом поле, смоделированные с правильным размером и структурой, но смещенные в пространстве относительно положения, принятого за истинное, могут дать очень плохие показатели поточечных метрик.

В терминах таблицы сопряженности под ложной тревогой понимаются случаи, когда событие прогнозируется, но не происходит, а промах – это наблюдаемое событие, которое не было спрогнозировано. Если объект немного смещен в пространстве, появятся две области с очень высокими значениями поточечных метрик – область с промахами и область с ложными тревогами, приводя к «двойному штрафу» поточечной метрики. Сильная пространственная изменчивость (например, как часто бывает с высокоразрешенными полями) приводит к большей вероятности наличия поточечных ошибок. При таких обстоятельствах «двойной штраф» может стать препятствием для оценки истинного качества прогноза (Gilleland et al., 2009).

В работе (Davis et al., 2006) впервые представлена систематизированная концепция феноменологического (*object-based*) метода, основанного на идентификации некоторых «объектов» в полях геофизических переменных и сравнении их атрибутов по данным моделирования/прогноза и фактическим данным. Примерами таких «объектов» могут являться области осадков, тропические циклоны, мезомасштабные когерентные структуры (например, полярные мезоциклоны, новоземельская бора) и т. д. Последние два явления представляют собой уникальные особенности мезомасштабной атмосферной циркуляции Арктики, которые не разрешены на моделях общей циркуляции с низким пространственным разрешением, что также повышает необходимость правильной оценки качества методов, повышающих пространственное разрешение сеточных данных.

Важным свойством этой метрики является гибкость по отношению к атрибутам, присущим разным «объектам». Например, для оценки прогноза осадков области дождя аппроксимируются в пространстве эллипсами, для которых определяется ряд параметров: площадь, центроид, направление оси, соотношение сторон, кривизна (Davis et al., 2006).

В обзоре (Gilleland et al., 2009) группа подобных мер объединяет метрики, основанные на особенностях (*features-based*), объектах (*object-based*) и идентификации ячеек (*cell-identification*) в полях атмосферных переменных. Эту группу метрик для простоты мы также будем называть феноменологическими.

Независимо от оцениваемой геофизической переменной, феноменологические метрики идентифицируют особенности, или объекты, применяя пороговое значение к полю или полям переменных. Размер, форма и средняя интенсивность по каждому объекту рассчитываются и сравниваются для каждого отдельного объекта в пределах прогнозируемого/смоделированного и фактического поля (Gilleland et al., 2009).

На наш взгляд, к положительным сторонам феноменологических метрик можно отнести то, что они напрямую определяют ошибки местоположения особенностей в полях переменных, предполагая, что прогноз достаточно похож на факт, чтобы соответствующие объекты могли быть сопоставлены.

Главный недостаток феноменологических метрик также отмечен в (Gilleland et al., 2009). Поля при использовании таких мер могут быть подвержены «контринтуитивному слиянию и сопоставлению». Например, при идентификации объектов по порогу интенсивности в один объект могут быть объединены области, которые находятся далеко друг от друга. Это объясняется тем, что разные эксперты, идентифицирующие объекты, в случае, если этот шаг не автоматизирован, вероятно, будут объединять области в объекты по-разному.

Основное различие между феноменологическими метриками и схожими метриками деформации поля заключается в том, что метрики деформации поля в первую очередь работают со всем полем геофизических переменных сразу без идентификации объектов и сравнений характеристик внутри только этих объектов.

По нашему мнению, феноменологические метрики представляют собой наиболее перспективное направление верификации в задачах атмосферы и океана, концентрируясь на особенностях соответствующих геофизических полей и избегая недостатков поточечных метрик.

Другие комплексные метрики

В научной литературе, посвященной оценке качества результатов различных моделей с атмосферными и океаническими данными, встречаются и другие метрики, не настолько широко распространенные, но также представляющие интерес.

Наравне со смещением 98-го перцентиля, описанным выше, у прогноза и факта также сравнивают параметр асимметрии – третий центральный момент распределения, являющийся показателем скошенности в направлении длинного «хвоста». Близкое значение асимметрии свидетельствует о правильности воспроизведения экстремальных значений величины, что важно, например, в атмосферном моделировании (Hertig et al., 2019).

Вероятностные прогнозы могут рассматриваться как бинарные переменные, принимающие значение «событие произошло», если прогнозируемая вероятность превышает заданный порог (Kharin, Zwiers, 2003). Как было сказано выше, под ложной тревогой понимаются случаи, когда событие спрогнозировано, но не происходит. При «попадании» спрогнозированное событие в действительности происходит. Для обработанного таким образом вероятностного прогноза можно построить кривую

ROC (*relative operating characteristic*), представляющую собой график частоты ложных тревог на оси *x* относительно частоты попаданий на оси *y*.

Нормализованная геометрическая площадь под кривой *ROC* называется *ROCSS* (*ROC skill score*) и часто используется для оценки качества вероятностных прогнозов (Buizza, Palmer, 1998; Mason, Graham, 1999; Kharin, Zwiers, 2003; Manzanas et al., 2014). Эта метрика принимает значения от 0 до 1. При этом значение 1 указывает на идеальную классификацию, значение 0.5 указывает на качество модели, эквивалентной случайному угадыванию, а значение 0 указывает на полностью неверную классификацию (Sun et al., 2024).

Несмотря на описанную ранее критику метрик качества, основанных на проверке экстремальных явлений (Lerch et al., 2017; Schultz et al., 2021), важно предупредить население об опасных метеорологических явлениях, которые могут повлечь за собой крупный ущерб и потери. Для таких экстремальных явлений необходимо создавать специализированные метрики (Casati et al., 2008).

В самом простом подходе экстремальное событие рассматривается как бинарная переменная с точки зрения его (не)наступления. Качество детерминистических прогнозов таких событий обычно оценивается с использованием традиционных категориальных оценок для бинарных событий (Casati et al., 2008). Тем не менее, такие метрики имеют тенденцию к неинформативным пределам 0 или 1 для все более редких событий (Stephenson et al., 2008).

В той же работе (Stephenson et al., 2008) предлагается более сложная метрика – оценка экстремальной зависимости (*EDS*), разработанная для оценки качества детерминированного прогнозирования редких бинарных событий. Одним из преимуществ *EDS* является отсутствие явной зависимости от выбора порога наступления экстремального явления (Casati et al., 2008).

Низкая частота экстремальных явлений создает определенные проблемы для оценки качества. Во-первых, редкость событий может привести к большой неопределенности выборки. Эта проблема частично решается путем объединения наблюдений и прогнозов в более крупных пространственно-временных областях, однако более крупные области также могут приводить к неоднородности и нестационарности (Casati et al., 2008). Во-вторых, редкость экстремальных явлений может привести к небольшому или нулевому количеству событий при разделении событий по классам. В-третьих, при небольших размерах выборки могут оказывать значительное влияние выбросы, искажающие проверку прогнозов экстремальных явлений (Casati et al., 2008).

Таким образом, для снижения ошибки оценки экстремальных явлений предлагается использовать метрики в подходе не детерминированных, а вероятностных прогнозов (Murphy, 1991; Casati et al., 2008).

Заключение

В настоящей статье мы представили обзор текущего состояния исследований метрик качества в задачах атмосферы и океана, кратко описав наиболее часто применяемые типы метрик качества и их преимущества и недостатки на основе имеющейся научной литературы. Мы представили наиболее общую систематизацию метрик качества и описали свойства метрик, применяемых к сеточным атмосферным и океаническим данным.

Наиболее широко в геофизических задачах, в частности, для увеличения пространственного разрешения геофизических данных, используются поточечные метрики. Такие метрики, чаще всего, применяются для непрерывных переменных. Наш анализ существующей на данный момент научной литературы показал значительное преобладание «большой тройки» классических метрик – средней арифметической ошибки, средней квадратической ошибки и коэффициента корреляции.

Мы отметили универсальность применения таких метрик к сеточным данным, при этом подчеркнув, что они, как и другие поточечные меры, обладают рядом существенных недостатков. Наиболее существенными препятствиями для повсеместного применения классических метрик мы полагаем непригодность к оценке качества воспроизведения внутренней пространственной корреляции полей и чувствительность к сдвигам неоднородностей. В ряде исследованных работ сделан более широкий вывод, с которым мы согласны лишь отчасти, о принципиальной неприменимости поточечных метрик в геофизических задачах. В таких рассуждениях мы поддерживаем то, что количественная оценка качества моделей представляет собой более сложную задачу.

На основе сформулированного нами тезиса о проблемах поточечных метрик мы расширили нашу работу с помощью описания различных комплексных метрик, сконцентрированных на более узких аспектах моделирования. Каждый из недостатков классических метрик позволил выделить направления разработки более сложных метрик качества.

Таким образом, отдельное внимание среди комплексных метрик мы посвятили мерам, учитывающим пространственную структуру и неоднородности геофизических полей, а также статистику данных. Мы также отметили существование окрестных метрик, рассматривающих данные с помощью скользящих окон; вероятностную оценку качества ансамблевых прогнозов и метрики для бинарных и категориальных переменных.

Мы подчеркнули противоречивость применения метрик, основанных на редких и экстремальных явлениях. С одной стороны, такие метрики важны для оценки того, насколько хорошо модель воспроизводит такие потенциально опасные явления. С другой стороны, мы призываем с осторожностью подходить к исключительному использованию таких мер, подвергающих всю модель «дилемме прогнозиста».

Самым многообещающим направлением развития комплексных метрик качества, на наш взгляд, является разработка феноменологических метрик на

основе алгоритмов идентификации «объектов» в пространственных геофизических полях.

Благодарности. Исследование поточечных мер качества выполнено в рамках Госзадания № FMWE-2022-0002. Обзор основных комплексных метрик выполнен при поддержке стратегического проекта «Исследовательское лидерство (от бакалавра до нобелевского лауреата)» программы «Приоритет-2030» Московского физико-технического института. Анализ феноменологических метрик и метрик, применимых к оценке редких и экстремальных событий, проведен в рамках гранта Российского научного фонда № 23-77-30001.

Список литературы

1. *Бундель А. Ю., Муравьев А. В., Ольховая Е. Д.* Обзор методов пространственной верификации и их применение для ансамблевых прогнозов // Гидрометеорологические исследования и прогнозы. 2021. № 4 (382). С. 30. https://doi.org/10.37162/2618-9631-2021-4-30-49.
2. *Калмыкова О. В.* Методика оценки смерчеопасности вблизи Черноморского побережья России и результаты ее испытаний // Результаты испытания новых и усовершенствованных технологий, моделей и методов гидрометеорологических прогнозов. 2021. №. 48. С. 42–61.
3. РД 52.27.284-91. Методические указания. Проведение производственных (оперативных) испытаний новых и усовершенствованных методов гидрометеорологических и гелиофизических прогнозов. Л.: Гидрометеоиздат, 1991. 150 с.
4. *Alessi M. J., DeGaetano A. T.* A comparison of statistical and dynamical downscaling methods for short-term weather forecasts in the US Northeast // Meteorological Applications. 2021. Vol. 28. No. 1. https://doi.org/10.1002/met.1976.
5. *Baño-Medina J., Manzanas R., Gutiérrez J. M.* On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections // Climate Dynamics. 2021. Vol. 57. No. 11. P. 2941–2951. https://doi.org/10.1007/s00382-021-05847-0.
6. *Barnston A. G., Thomas J. L.* Rainfall measurement accuracy in FACE: A comparison of gage and radar rainfalls // Journal of Applied Meteorology and Climatology. 1983. Vol. 22. No. 12. P. 2038–2052. https://doi.org/10.1175/1520-0450(1983)022%3C2038:RMAIFA%3E2.0.CO;2.
7. *Barnston A. G.* Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score // Weather and Forecasting. 1992. Vol. 7. No. 4. P. 699–709. https://doi.org/10.1175/1520-0434(1992)007%3C0699:CATCRA%3E2.0.CO;2.
8. *Barsugli J. J., Guentchev G., Horton R. M., Wood A., Mearns L. O., Liang X. Z., Winkler J. A., Dixon K., Hayhoe K., Rood R. B., Goddard L., Ray A., Buja L., Ammann C.* The practitioner's dilemma: How to assess the credibility of downscaled climate projections // Eos, Transactions American Geophysical Union. 2013. Vol. 94. No. 46. P. 424–425. https://doi.org/10.1002/2013EO460005.
9. *Böker B., Laux P., Olschewski P., Kunstmann H.* Added value of an atmospheric circulation pattern-based statistical downscaling approach for daily precipitation distributions in complex terrain // International Journal of Climatology. 2023. Vol. 43. No. 11. P. 5130–5153. https://doi.org/10.1002/joc.8136.
10. *Briley L., Kelly R., Blackmer E. D., Troncoso A. V., Rood R. B., Andresen, J., Lemos M. C.* Increasing the usability of climate models through the use of consumer-report-style resources for decision-making // Bulletin of the American Meteorological Society. 2020. Vol. 101. No. 10. P. E1709–E1717. https://doi.org/10.1175/BAMS-D-19-0099.1.

11. *Bröcker J.* Evaluating raw ensembles with the continuous ranked probability score // Quarterly Journal of the Royal Meteorological Society. 2012. Vol. 138. No. 667. P. 1611–1617. https://doi.org/10.1002/qj.1891.
12. *Buizza R., Palmer T. N.* Impact of ensemble size on ensemble prediction // Monthly Weather Review. 1998. Vol. 126. No. 9. P. 2503–2518. https://doi.org/10.1175/1520-0493(1998)126%3C2503:IOESOE%3E2.0.CO;2.
13. *Buschow S., Pidstrigach J., Friederichs P.* Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0. 1.0) // Geoscientific Model Development. 2019. Vol. 12. No. 8. P. 3401–3418. https://doi.org/10.5194/gmd-12-3401-2019.
14. *Carroll R. J., Ruppert D., Stefanski L. A., Crainiceanu C. M.* Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC, 2006.
15. *Casati B., Wilson L. J., Stephenson D. B., Nurmi P., Ghelli A., Pocernich M., Damrath U., Ebert E. E., Brown B. G., Mason S.* Forecast verification: current status and future directions // Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling. 2008. Vol. 15. No. 1. P. 3–18. https://doi.org/10.1002/met.52.
16. *Chai T., Draxler R. R.* Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature // Geoscientific model development. 2014. Vol. 7. No. 3. P. 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.
17. *Cho D., Im J., Jung S.* A new statistical downscaling approach for short-term forecasting of summer air temperatures through a fusion of deep learning and spatial interpolation // Quarterly Journal of the Royal Meteorological Society. 2024. Vol. 150. No. 760. P. 1222–1242. https://doi.org/10.1002/qj.4643.
18. *Collier N., Hoffman F. M., Lawrence D. M., Keppel-Aleks G., Koven C. D., Riley W. J., Mu M., Randerson J. T.* The international land model benchmarking (ILAMB) system: Design, theory, and implementation // Journal of Advances in Modeling Earth Systems. 2018. Vol. 10. No. 11. P. 2731–2754. https://doi.org/10.1029/2018MS001354.
19. *Davis C., Brown B., Bullock R.* Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas // Monthly Weather Review. 2006. Vol. 134. No. 7. P. 1772–1784. https://doi.org/10.1175/MWR3145.1.
20. *De Haan L. L., Kanamitsu M., De Sales F., Sun, L.* An evaluation of the seasonal added value of downscaling over the United States using new verification measures // Theoretical and applied climatology. 2015. Vol. 122. P. 47–57. https://doi.org/10.1007/s00704-014-1278-9.
21. *Dosselmann R., Yang X. D.* A comprehensive assessment of the structural similarity index // Signal, Image and Video Processing. 2011. Vol. 5. P. 81–91. https://doi.org/10.1007/s11760-009-0144-1.
22. *Ebert E. E., McBride J. L.* Verification of precipitation in weather systems: Determination of systematic errors // Journal of hydrology. 2000. Vol. 239. No. 1–4. P. 179–202. https://doi.org/10.1016/S0022-1694(00)00343-7.
23. *Ebert E. E.* Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework // Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling. 2008. Vol. 15. No. 1. P. 51–64. https://doi.org/10.1002/met.25.
24. *Ebert E., Wilson L., Weigel A., Mittermaier M., Nurmi P., Gill P., Göber M., Joslyn S., Brown B., Fowler T., Watkins A.* Progress and challenges in forecast verification // Meteorological Applications. 2013. Vol. 20. No. 2. P. 130–139. https://doi.org/10.1002/met.1392.
25. *Ebert P., Milne P.* Methodological and conceptual challenges in rare and severe event forecast-verification // Natural Hazards and Earth System Sciences Discussions. 2022. Vol. 20. P. 539–557. https://doi.org/10.5194/nhess-22-539-2022.
26. *Ebtehaj A. M., Foufoula-Georgiou E., Lerman G.* Sparse regularization for precipitation downscaling // Journal of Geophysical Research: Atmospheres. 2012. Vol. 117. No. D8. https://doi.org/10.1029/2011JD017057.

27. *Entekhabi D., Reichle R. H., Koster R. D., Crow W. T.* Performance metrics for soil moisture retrievals and application requirements // Journal of Hydrometeorology. 2010. Vol. 11. No. 3. P. 832–840. https://doi.org/10.1175/2010JHM1223.1.
28. *Faggian N., Roux B., Steinle P., Ebert B.* Fast calculation of the fractions skill score // Mausam. 2015. Vol. 66. No. 3. P. 457–466. http://dx.doi.org/10.54302/mausam.v66i3.555.
29. *Fisher S. R. A.* Statistical Methods for Research Workers. 13th ed. Hafner Publishing Co., 1958. 356 p.
30. *Geiss A., Silva S. J., Hardin J. C.* Downscaling atmospheric chemistry simulations with physically consistent deep learning // Geoscientific Model Development. 2022. Vol. 15. No. 17. P. 6677–6694. https://doi.org/10.5194/gmd-15-6677-2022.
31. *Gentile E. S., Gray S. L., Lewis H. W.* The sensitivity of probabilistic convective-scale forecasts of an extratropical cyclone to atmosphere–ocean–wave coupling // Quarterly Journal of the Royal Meteorological Society. 2022. Vol. 148. No. 743. P. 685–710. https://doi.org/10.1002/qj.4225.
32. *Gilabert M. A., Moreno A., Maselli F., Martinez B., Chiesi M., Sanchez-Ruiz S., Garcia-Haro F. J., Perez-Hoyos A., Campos-Taberner M., Perez-Priego O., Serrano-Ortiz P., Carrara A.* Daily GPP estimates in Mediterranean ecosystems by combining remote sensing and meteorological data // ISPRS Journal of Photogrammetry and Remote Sensing. 2015. Vol. 102. P. 184–197. https://doi.org/10.1016/j.isprsjprs.2015.01.017.
33. *Gilleland E., Ahijevych D., Brown B. G., Casati B., Ebert E. E.* Intercomparison of Spatial Forecast Verification Methods // Weather Forecasting. 2009. Vol. 24. P. 1416–1430. https://doi.org/10.1175/2009WAF2222269.1.
34. *Glawion L., Polz J., Kunstmann H., Fersch B., Chwala C.* spateGAN: spatio-temporal downscaling of rainfall fields using a cGAN approach // Earth and Space Science. 2023. Vol. 10. No. 10. https://doi.org/10.1029/2023EA002906.
35. *Gong W., Gupta H. V., Yang D., Sricharan K., Hero III A. O.* Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach // Water resources research. 2013. Vol. 49. No. 4. P. 2253–2273. https://doi.org/10.1002/wrcr.20161.
36. *Gupta H. V., Kling H., Yilmaz K. K., Martinez G. F.* Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling // Journal of hydrology. 2009. Vol. 377. No. 1–2. P. 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
37. *Habib E., Krajewski W. F., Ciach G. J.* Estimation of rainfall interstation correlation // Journal of Hydrometeorology. 2001. Vol. 2. No. 6. P. 621–629. https://doi.org/10.1175/1525-7541(2001)002%3C0621:EORIC%3E2.0.CO;2.
38. *Harris L., McRae A. T., Chantry M., Dueben P. D., Palmer T. N.* A generative deep learning approach to stochastic downscaling of precipitation forecasts // Journal of Advances in Modeling Earth Systems. 2022. Vol. 14. No. 10. https://doi.org/10.1029/2022MS003120.
39. *Hersbach H.* Decomposition of the continuous ranked probability score for ensemble prediction systems // Weather and Forecasting. 2000. Vol. 15. No. 5. P. 559–570. https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2.
40. *Hertig E., Maraun D., Bartholy J., Pongracz R., Vrac M., Mares I., Gutiérrez J. M., Wibig J., Casanueva A., Soares P. M. M.* Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE // International Journal of Climatology. 2019. Vol. 39. No. 9. P. 3846–3867. https://doi.org/10.1002/joc.5469.
41. *Hewitson B. C., Crane R. G.* Climate downscaling: techniques and application // Climate Research. 1996. Vol. 7. No. 2. P. 85–95. https://doi.org/10.3354/cr007085.
42. *Hodson T. O.* Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not // Geoscientific Model Development Discussions. 2022. Vol. 15. No. 14. P. 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022.

43. *Jagannathan K., Jones A. D., Ray I.* The making of a metric: Coproducing decision-relevant climate science // Bulletin of the American Meteorological Society. 2020. Vol. 102. No. 8. P. 1–33. https://doi.org/10.1175/BAMS-D-19-0296.1.
44. *Jin H., Jiang W., Chen M., Li M., Bakar K. S., Shao Q.* Downscaling long lead time daily rainfall ensemble forecasts through deep learning // Stochastic Environmental Research and Risk Assessment. 2023. Vol. 37. No. 8. P. 3185–3203. https://doi.org/10.1007/s00477-023-02444-x.
45. *Jolliffe I. T., Stephenson D.* B. Forecast verification: a practitioner's guide in atmospheric science. John Wiley & Sons, 2012.
46. *Kharin V. V., Zwiers F. W.* On the ROC score of probability forecasts // Journal of Climate. 2003. Vol. 16. No. 24. P. 4145–4150. https://doi.org/10.1175/1520-0442(2003)016%3C4145:OTRSOP%3E2.0.CO;2.
47. *Kikumoto H., Ooka R., Arima Y., Yamanaka T.* Study on the future weather data considering the global and local climate change for building energy simulation // Sustainable Cities and Society. 2015. Vol. 14. P. 404–413. https://doi.org/10.1016/j.scs.2014.08.007.
48. *Kling H., Fuchs M., Paulin M.* Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios // Journal of hydrology. 2012. Vol. 424. P. 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011.
49. *Knoben W. J. M., Freer J. E., Woods R. A.* Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores // Hydrology and Earth System Sciences. 2019. Vol. 23. No. 10. P. 4323–4331. https://doi.org/10.5194/hess-23-4323-2019.
50. *Lerch S., Thorarinsdottir T. L., Ravazzolo F., Gneiting T.* Forecaster's dilemma: extreme events and forecast evaluation // Statistical Science. 2017. Vol. 32. No. 1. P. 106–127. https://www.jstor.org/stable/26408123.
51. *Manzanas R., Frías M. D., Cofiño A. S., Gutiérrez J. M.* Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill // Journal of Geophysical Research: Atmospheres. 2014. Vol. 119. No. 4. P. 1708–1719. https://doi.org/10.1002/2013JD020680.
52. *Maina F. Z., Siirila-Woodburn E. R., Vahmani P.* Sensitivity of meteorological-forcing resolution on hydrologic variables // Hydrology and Earth System Sciences. 2020. Vol. 24. No. 7. P. 3451–3474. https://doi.org/10.5194/hess-24-3451-2020.
53. *Mason S. J., Graham N. E.* Conditional probabilities, relative operating characteristics, and relative operating levels // Weather and Forecasting. 1999. Vol. 14. No. 5. P. 713–725. https://doi.org/10.1175/1520-0434(1999)014%3C0713:CPROCA%3E2.0.CO;2.
54. *McKeen S. A., Wilczak J., Grell G., Djalalova I., Peckham S., Hsie E., Gong W., Bouchet V., Menard S., Moffet R., McHenry J., McQueen J., Tang Y., Carmichael G. R., Pagowski M., Chan A., Dye T., Frost G., Lee P., Mathur R.* Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004 // Journal of Geophysical Research: Atmospheres. 2005. Vol. 110. No. D21. https://doi.org/10.1029/2005JD005858.
55. *Mittermaier M., Roberts N., Thompson S. A.* A long-term assessment of precipitation forecast skill using the fractions skill score // Meteorological Applications. 2013. Vol. 20. No. 2. P. 176–186. https://doi.org/10.1002/met.296.
56. *Moss R. H., Avery S., Baja K., Burkett M., Chischilly A. M., Dell J., Fleming P. A., Geil K., Jacobs K., Jones A., Knowlton K., Koh J., Lemos M. C., Melillo J., Pandya R., Richmond T. C., Scarlett L., Snyder J., Stults M., Waple A. M., Whitehead J., Zarrilli D., Ayyub B. M., Fox J., Ganguly A., Joppa L., Julius S., Kirshen P., Kreutter R., McGovern A., Meyer R., Neumann J., Solecki W., Smith J., Tissot P., Yohe G., Zimmerman R.* Evaluating knowledge to support climate action: A framework for sustained assessment. Report of an independent advisory committee on applied climate assessment // Weather, Climate, and Society. 2019. Vol. 11. No. 3. P. 465–487. https://doi.org/10.1175/WCAS-D-18-0134.1.

57. *Murphy A. H.* Skill scores based on the mean square error and their relationships to the correlation coefficient // Monthly weather review. 1988. Vol. 116. No. 12. P. 2417–2424. https://doi.org/10.1175/1520-0493(1988)116%3C2417:SSBOTM%3E2.0.CO;2.
58. *Murphy A. H., Brown B. G., Chen Y. S.* Diagnostic verification of temperature forecasts // Weather and Forecasting. 1989. Vol. 4. No. 4. P. 485–501. https://doi.org/10.1175/1520-0434(1989)004%3C0485:DVOTF%3E2.0.CO;2.
59. *Murphy A. H.* Probabilities, odds, and forecasts of rare events // Weather and forecasting. 1991. Vol. 6. No. 2. P. 302–307. https://doi.org/10.1175/1520-0434(1991)006%3C0302:POAFOR%3E2.0.CO;2.
60. *Murphy A. H.* What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting // Weather Forecasting. 1993. Vol. 8. No. 2. P. 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.
61. *Nearing G. S., Gupta H. V.* The quantity and quality of information in hydrologic models // Water Resources Research. 2015. Vol. 51. No. 1. P. 524–538. https://doi.org/10.1002/2014WR015895.
62. *O'Neill A. C., Erikson L. H., Barnard P. L.* Downscaling wind and wavefields for 21st century coastal flood hazard projections in a region of complex terrain // Earth and Space Science. 2017. Vol. 4. No. 5. P. 314–334. https://doi.org/10.1002/2016EA000193.
63. *Pool S., Vis M., Seibert J.* Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency // Hydrological Sciences Journal. 2018. Vol. 63. No. 13–14. P. 1941–1953. https://doi.org/10.1080/02626667.2018.1552002.
64. *Reed K. A., Goldenson N., Grotjahn R., Gutowski W. J., Jagannathan K., Jones A. D., Leung L. R., McGinnis S. A., Pryor S. C., Srivastava A. K., Ullrich P. A., Zarzycki C. M.* Metrics as tools for bridging climate science and applications // Wiley Interdisciplinary Reviews: Climate Change. 2022. Vol. 13. No. 6. P. e799. https://doi.org/10.1002/wcc.799.
65. *Roberts N. M.* Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model // Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling. 2008. Vol. 15. No. 1. P. 163–169. https://doi.org/10.1002/met.57.
66. *Roberts N. M., Lean H. W.* Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events // Monthly Weather Review. 2008. Vol. 136. No. 1. P. 78–97. https://doi.org/10.1175/2007MWR2123.1.
67. *Robertson C., Long J. A., Nathoo F. S., Nelson T. A., Plouffe C. C.* Assessing quality of spatial models using the structural similarity index and posterior predictive checks // Geographical Analysis. 2014. Vol. 46. No. 1. P. 53–74. https://doi.org/10.1111/gean.12028.
68. *Savage N. H., Agnew P., Davis L. S., Ordóñez C., Thorpe R., Johnson C. E., O'Connor F. M., Dalvi M.* Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation // Geoscientific Model Development. 2013. Vol. 6. No. 2. P. 353–372. https://doi.org/10.5194/gmd-6-353-2013.
69. *Schoof J. T.* Statistical downscaling in climatology // Geography Compass. 2013. Vol. 7. No. 4. P. 249–265. https://doi.org/10.1111/gec3.12036.
70. *Schultz M. G., Betancourt C., Gong B., Kleinert F., Langguth M., Leufen L. H., Mozaffari A., Stadtler S.* Can deep learning beat numerical weather prediction? // Philosophical Transactions of the Royal Society A. 2021. Vol. 379. No. 2194. https://doi.org/10.1098/rsta.2020.0097.
71. *Shahriar S. A., Siddique M. A. M., Rahman S. M. A.* Climate change projection using statistical downscaling model over Chittagong Division, Bangladesh // Meteorology and Atmospheric Physics. 2021. Vol. 133. No. 4. P. 1409–1427. https://doi.org/10.1007/s00703-021-00817-x.
72. *Skok G.* Analysis of fraction skill score properties for a displaced rainband in a rectangular domain // Meteorological Applications. 2015. Vol. 22. No. 3. P. 477–484. https://doi.org/10.1002/met.1478.

73. *Skok G., Roberts N.* Estimating the displacement in precipitation forecasts using the Fractions Skill Score // Quarterly Journal of the Royal Meteorological Society. 2018. Vol. 144. No. 711. P. 414–425. https://doi.org/10.1002/qj.3212.
74. *Song L., Xu C., Long Y., Lei X., Suo N., Cao L.* Performance of seven gridded precipitation products over arid central Asia and subregions // Remote Sensing. 2022. Vol. 14. No. 23. P. 6039. https://doi.org/10.3390/rs14236039.
75. *Stephenson D. B., Casati B., Ferro C. A. T., Wilson C. A.* The extreme dependency score: A non-vanishing measure for forecasts of rare events // Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling. 2008. Vol. 15. No. 1. P. 41–50. https://doi.org/10.1002/met.53.
76. *Sun Y., Deng K., Ren K., Liu J., Deng C., Jin Y.* Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review // ISPRS Journal of Photogrammetry and Remote Sensing. 2024. Vol. 208. P. 14–38. https://doi.org/10.1016/j.isprsjprs.2023.12.011.
77. *Taylor K. E.* Summarizing multiple aspects of model performance in a single diagram // Journal of Geophysical Research: Atmospheres. 2001. Vol. 106. P. 7183–7192. https://doi.org/10.1029/2000JD900719.
78. *Tian Y., Nearing G. S., Peters-Lidard C. D., Harrison K. W., Tang L.* Performance metrics, error modeling, and uncertainty quantification // Monthly Weather Review. 2016. Vol. 144. No. 2. P. 607–613. https://doi.org/10.1175/MWR-D-15-0087.1.
79. *Vaughan A., Tebbutt W., Hosking J. S., Turner R. E.* Convolutional conditional neural processes for local climate downscaling // Geoscientific Model Development. 2022. Vol. 15. No. 1. P. 251–268. https://doi.org/10.5194/gmd-15-251-2022.
80. *Vrugt J. A., Sadegh M.* Toward diagnostic model calibration and evaluation: Approximate Bayesian computation // Water Resources Research. 2013. Vol. 49. No. 7. P. 4335–4345. https://doi.org/10.1002/wrcr.20354.
81. *Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.* Image quality assessment: from error visibility to structural similarity // IEEE transactions on image processing. 2004. Vol. 13. No. 4. P. 600–612. https://doi.org/10.1109/TIP.2003.819861.
82. *Wehner M. F.* Very extreme seasonal precipitation in the NARCCAP ensemble: Model performance and projections // Climate Dynamics. 2013. Vol. 40. P. 59–80. https://doi.org/10.1007/s00382-012-1393-1.
83. *Weijs S. V., Schoups G., Van De Giesen N.* Why hydrological predictions should be evaluated using information theory // Hydrology and Earth System Sciences. 2010. Vol. 14. No. 12. P. 2545–2558. https://doi.org/10.5194/hess-14-2545-2010.
84. *Weniger M., Kapp F., Friederichs P.* Spatial verification using wavelet transforms: a review // Quarterly Journal of the Royal Meteorological Society. 2017. Vol. 143. No. 702. P. 120–136. https://doi.org/10.1002/qj.2881.
85. *Wilby R. L., Wigley T. M. L.* Downscaling general circulation model output: a review of methods and limitations // Progress in physical geography. 1997. Vol. 21. No. 4. P. 530–548. https://doi.org/10.1177/030913339702100403.
86. *Wilks D. S.* Statistical methods in the atmospheric sciences. Academic press, 2011. 676 p.
87. *Willmott C. J., Matsuura K.* Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance // Climate research. 2005. Vol. 30. No. 1. P. 79–82. https://doi.org/10.3354/cr030079.
88. *Willmott C. J., Matsuura K., Robeson S. M.* Ambiguities inherent in sums-of-squares-based error statistics // Atmospheric Environment. 2009. Vol. 43. No. 3. P. 749–752. https://doi.org/10.1016/j.atmosenv.2008.10.005.
89. *Yeggina S., Teegavarapu R. S. V., Muddu S.* Evaluation and bias corrections of gridded precipitation data for hydrologic modelling support in Kabini River basin, India // Theoretical and Applied Climatology. 2020. Vol. 140. P. 1495–1513. https://doi.org/10.1007/s00704-020-03175-7.

90. *Zandler H., Senftl T., Vanselow K. A.* Reanalysis datasets outperform other gridded climate products in vegetation change analysis in peripheral conservation areas of Central Asia // Scientific Reports. 2020. Vol. 10. No. 1. https://doi.org/10.1038/s41598-020-79480-y.
91. *Zhong X., Du F., Chen L., Wang Z., Li H.* Investigating transformer-based models for spatial downscaling and correcting biases of near-surface temperature and wind-speed forecasts // Quarterly Journal of the Royal Meteorological Society. 2024. Vol. 150. No. 758. P. 275–289. https://doi.org/10.1002/qj.4596.

Статья поступила в редакцию 03.06.2024, одобрена к печати 10.12.2024.
Для цитирования: *Резвов В. Ю., Криницкий М. А., Тилинина Н. Д.* Поточечные и комплексные меры качества в исследованиях атмосферы и океана: обзор методов и подходов // Океанологические исследования. 2024. № 52 (4). С. 193–223. https://doi.org/10.29006/1564-2291.JOR-2024.52(4).10.

POINT-BY-POINT AND COMPLEX QUALITY METRICS IN ATMOSPHERE AND OCEAN RESEARCH: REVIEW OF METHODS AND APPROACHES

V. Yu. Rezvov[1,2], M. A. Krinitskiy[1,2], N. D. Tilinina[1]

[1] Shirshov Institute of Oceanology, Russian Academy of Sciences,
36, Nakhimovskiy prospekt, Moscow, 117997, Russia;
[2] Moscow Institute of Physics and Technology (National Research University),
9, Institutskiy pereulok, Dolgoprudny, 141701, Russia,
e-mail: rezvov.vyu@phystech.su

In the oceanic and atmospheric sciences, various general quantitative indicators, or quality metrics, describe the quality of the various modeling products, including numerical weather prediction, statistical correction, and downscaling. Metrics provide the level of accuracy of model processes reproduction and allow for comparison of models by estimating the uncertainty of their results. This paper presents a classification of the most frequently encountered quality metrics in the scientific literature. Examples are given for each group of quality metrics. In addition to assessing traditional point-by-point metrics, complex metrics that consider various aspects of modeling results are studied. Such specific metrics have an emphasis on the spatial structure, internal correlations, and heterogeneity of the predicted variable fields, ensemble forecasts etc. Special attention in this paper is also devoted to the object-oriented metrics or metrics based for rare and extreme events.

Keywords: quality metrics, grid data, downscaling, verification, point-by-point metrics, complex metrics, ensemble forecasting, object-oriented metrics

Acknowledgement: The study of point-by-point quality measures was carried out within the framework of State Assignment No. FMWE-2022-0002. The review of the basic complex metrics was provided with the support of the strategic project “Research Leadership (from Bachelor to Nobel Laureate)” of the Priority-2030 program of the Moscow Institute of Physics and Technology. The analysis of object-oriented metrics and metrics applicable to the rare and extreme events was carried out within the framework of the grant of the Russian Science Foundation No. 23-77-30001.

References

1. Alessi, M. J. and A. T. DeGaetano, 2021: A comparison of statistical and dynamical downscaling methods for short-term weather forecasts in the US Northeast. *Meteorological Applications*, **28** (1), https://doi.org/10.1002/met.1976.
2. Baño-Medina, J., R. Manzanas, and J. M. Gutiérrez, 2021: On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, **57** (11), 2941–2951, https://doi.org/10.1007/s00382-021-05847-0.
3. Barnston, A. G. and J. L. Thomas, 1983: Rainfall measurement accuracy in FACE: A comparison of gage and radar rainfalls. *Journal of Applied Meteorology and Climatology*, **22** (12), 2038–2052, https://doi.org/10.1175/1520-0450(1983)022%3C2038:RMAIFA%3E2.0.CO;2.
4. Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, **7** (4), 699–709, https://doi.org/10.1175/1520-0434(1992)007%3C0699:CATCRA%3E2.0.CO;2.
5. Barsugli, J. J., G. Guentchev, R. M. Horton, A. Wood, L. O. Mearns, X. Z. Liang, J. A. Winkler, K. Dixon, K. Hayhoe, R. B. Rood, L. Goddard, A. Ray, L. Buja, and C. Ammann, 2013: The practitioner's dilemma: How to assess the credibility of downscaled climate projections. *Eos, Transactions American Geophysical Union*, **94** (46), 424–425, https://doi.org/10.1002/2013EO460005.
6. Böker, B., P. Laux, P. Olschewski, and H. Kunstmann, 2023: Added value of an atmospheric circulation pattern-based statistical downscaling approach for daily precipitation distributions in complex terrain. *International Journal of Climatology*, **43** (11), 5130–5153, https://doi.org/10.1002/joc.8136.
7. Briley, L., R. Kelly, E. D. Blackmer, A. V. Troncoso, R. B. Rood, J. Andresen, and M. C. Lemos, 2020: Increasing the usability of climate models through the use of consumer-report-style resources for decision-making. *Bulletin of the American Meteorological Society*, **101** (10), E1709–E1717, https://doi.org/10.1175/BAMS-D-19-0099.1.
8. Bröcker, J., 2012: Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, **138** (667), 1611–1617, https://doi.org/10.1002/qj.1891.
9. Buizza, R. and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Monthly Weather Review*, **126** (9), 2503–2518, https://doi.org/10.1175/1520-0493(1998)126%3C2503:IOESOE%3E2.0.CO;2.
10. Bundel, A. Yu., A. V. Muravev, and E. D. Olhovaya, 2021: Obzor metodov prostranstvennoj verifikacii i ih primenenie dlya ansamblevyh prognozov (Review of spatial verification methods and their application to ensemble forecasts). *Gidrometeorologicheskie issledovaniya i prognozy*, **4** (382), 30, https://doi.org/10.37162/2618-9631-2021-4-30-49.
11. Buschow, S., J. Pidstrigach, and P. Friederichs, 2019: Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0. 1.0). *Geoscientific Model Development*, **12** (8), 3401–3418, https://doi.org/10.5194/gmd-12-3401-2019.
12. Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, 2006: *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.
13. Casati, B., L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason, 2008: Forecast verification: current status and future directions. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, **15** (1), 3–18, https://doi.org/10.1002/met.52.
14. Chai, T., and R. R. Draxler, 2014: Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, **7** (3), 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014.

15. Cho, D., J. Im, and S. Jung, 2024: A new statistical downscaling approach for short-term forecasting of summer air temperatures through a fusion of deep learning and spatial interpolation. *Quarterly Journal of the Royal Meteorological Society*, **150** (760), 1222–1242, https://doi.org/10.1002/qj.4643.
16. Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson, 2018: The international land model benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, **10** (11), 2731–2754, https://doi.org/10.1029/2018MS001354.
17. Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134** (7), 1772–1784, https://doi.org/10.1175/MWR3145.1.
18. De Haan, L. L., M. Kanamitsu, F. De Sales, and L. Sun, 2015: An evaluation of the seasonal added value of downscaling over the United States using new verification measures. *Theoretical and applied climatology*, **122**, 47–57, https://doi.org/10.1007/s00704-014-1278-9.
19. Dosselmann, R. and X. D. Yang, 2011: A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, **5**, 81–91, https://doi.org/10.1007/s11760-009-0144-1.
20. Ebert, E. E. and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *Journal of hydrology*, **239** (1–4), 179–202, https://doi.org/10.1016/S0022-1694(00)00343-7.
21. Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, **15** (1), 51–64, https://doi.org/10.1002/met.25.
22. Ebert, E., L. Wilson, A. Weigel, M. Mittermaier, P. Nurmi, P. Gill, M. Göber, S. Joslyn, B. Brown, T. Fowler, and A. Watkins, 2013: Progress and challenges in forecast verification. *Meteorological Applications*, **20** (2), 130–139, https://doi.org/10.1002/met.1392.
23. Ebert, P. and P. Milne, 2022: Methodological and conceptual challenges in rare and severe event forecast-verification. *Natural Hazards and Earth System Sciences Discussions*, **20**, 539–557, https://doi.org/10.5194/nhess-22-539-2022.
24. Ebtehaj, A. M., E. Foufoula-Georgiou, and G. Lerman, 2012: Sparse regularization for precipitation downscaling. *Journal of Geophysical Research: Atmospheres*, **117** (D8), https://doi.org/10.1029/2011JD017057.
25. Entekhabi, D., R. H. Reichle, R. D. Koster, and W. T. Crow, 2010: Performance metrics for soil moisture retrievals and application requirements. *Journal of Hydrometeorology*, **11** (3), 832–840, https://doi.org/10.1175/2010JHM1223.1.
26. Faggian, N., B. Roux, P. Steinle, and B. Ebert, 2015: Fast calculation of the fractions skill score. *Mausam*, **66** (3), 457–466, http://dx.doi.org/10.54302/mausam.v66i3.555.
27. Fisher, S. R. A., 1958: *Statistical Methods for Research Workers*. 13th ed. Hafner Publishing Co., 356 p.
28. Geiss, A., S. J. Silva, and J. C. Hardin, 2022: Downscaling atmospheric chemistry simulations with physically consistent deep learning. *Geoscientific Model Development*, **15** (17), 6677–6694, https://doi.org/10.5194/gmd-15-6677-2022.
29. Gentile, E. S., S. L. Gray, and H. W. Lewis, 2022: The sensitivity of probabilistic convective-scale forecasts of an extratropical cyclone to atmosphere–ocean–wave coupling. *Quarterly Journal of the Royal Meteorological Society*, **148** (743), 685–710, https://doi.org/10.1002/qj.4225.
30. Gilabert, M. A., A. Moreno, F. Maselli, B. Martinez, M. Chiesi, S. Sanchez-Ruiz, F. J. Garcia-Haro, A. Perez-Hoyos, M. Campos-Taberner, O. Perez-Priego, P. Serrano-Ortiz, and A. Carrara, 2015: Daily GPP estimates in Mediterranean ecosystems by combining remote sensing and meteorological data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **102,** 184–197, https://doi.org/10.1016/j.isprsjprs.2015.01.017.

31. Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. *Weather Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.
32. Glawion, L., J. Polz, H. Kunstmann, B. Fersch, and C. Chwala, 2023: spateGAN: spatio-temporal downscaling of rainfall fields using a cGAN approach. *Earth and Space Science*, **10** (10), https://doi.org/10.1029/2023EA002906.
33. Gong, W., H. V. Gupta, D. Yang, K. Sricharan, and A. O. Hero III, 2013: Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water resources research*, **49** (4), 2253–2273, https://doi.org/10.1002/wrcr.20161.
34. Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, **377** (1–2), 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003.
35. Habib, E., W. F. Krajewski, and G. J. Ciach, 2001: Estimation of rainfall interstation correlation. *Journal of Hydrometeorology*, **2** (6), 621–629, https://doi.org/10.1175/1525-7541(2001)002%3C0621:EORIC%3E2.0.CO;2.
36. Harris, L., A. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, 2022: A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, **14** (10), https://doi.org/10.1029/2022MS003120.
37. Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15** (5), 559–570, https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2.
38. Hertig, E., D. Maraun, J. Bartholy, R. Pongracz, M. Vrac, I. Mares, J. M. Gutiérrez, J. Wibig, A. Casanueva, and P. M. M. Soares, 2019: Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE. *International Journal of Climatology*, **39** (9), 3846–3867, https://doi.org/10.1002/joc.5469.
39. Hewitson, B. C. and R. G. Crane, 1996: Climate downscaling: techniques and application. *Climate Research*, **7** (2), 85–95, https://doi.org/10.3354/cr007085.
40. Hodson, T. O., 2022: Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, **15** (14), 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022.
41. Jagannathan, K., A. D. Jones, and I. Ray, 2020: The making of a metric: Coproducing decision-relevant climate science. *Bulletin of the American Meteorological Society*, **102** (8), 1–33, https://doi.org/10.1175/BAMS-D-19-0296.1.
42. Jin, H., W. Jiang, M. Chen, M. Li, K. S. Bakar, and Q. Shao, 2023: Downscaling long lead time daily rainfall ensemble forecasts through deep learning. *Stochastic Environmental Research and Risk Assessment*, **37** (8), 3185–3203, https://doi.org/10.1007/s00477-023-02444-x.
43. Jolliffe, I. T. and D. B. Stephenson, 2012: F*orecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.
44. Kalmykova, O. V., 2021: Metodika ocenki smercheopasnosti vblizi Chernomorskogo poberezh'ya Rossii i rezul'taty ee ispytanij (Methodology for assessing tornado hazard near the Black Sea coast of Russia and the results of its tests). *Rezul'taty ispytaniya novyh i usovershenstvovannyh tekhnologij, modelej i metodov gidrometeorologicheskih prognozov*, **48**, 42–61.
45. Kharin, V. V. and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *Journal of Climate*, **16** (24), 4145–4150, https://doi.org/10.1175/1520-0442(2003)016%3C4145:OTRSOP%3E2.0.CO;2.
46. Kikumoto, H., R. Ooka, Y. Arima, and T. Yamanaka, 2015: Study on the future weather data considering the global and local climate change for building energy simulation. *Sustainable Cities and Society*, **14**, 404–413, https://doi.org/10.1016/j.scs.2014.08.007.

47. Kling, H., M. Fuchs, and M. Paulin, 2012: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of hydrology*, **424**, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011.
48. Knoben, W. J. M., J. E. Freer, and R. A. Woods, 2019: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, **23** (10), 4323–4331, https://doi.org/10.5194/hess-23-4323-2019.
49. Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, **32** (1), 106–127, https://www.jstor.org/stable/26408123.
50. Manzanas, R., M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez, 2014: Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *Journal of Geophysical Research: Atmospheres*, **119** (4), 1708–1719, https://doi.org/10.1002/2013JD020680.
51. Maina, F. Z., E. R. Siirila-Woodburn, and P. Vahmani, 2020: Sensitivity of meteorological-forcing resolution on hydrologic variables. *Hydrology and Earth System Sciences*, **24** (7), 3451–3474, https://doi.org/10.5194/hess-24-3451-2020.
52. Mason, S. J. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14** (5), 713–725, https://doi.org/10.1175/1520-0434(1999)014%3C0713:CPROCA%3E2.0.CO;2.
53. McKeen, S. A., J. Wilczak, G. Grell, I. Djalalova, S. Peckham, E. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, and R. Mathur, 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research: Atmospheres*, **110** (D21), https://doi.org/10.1029/2005JD005858.
54. Mittermaier, M., N. Roberts, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteorological Applications*, **20** (2), 176–186, https://doi.org/10.1002/met.296.
55. Moss, R. H., S. Avery, K. Baja, M. Burkett, A. M. Chischilly, J. Dell, P. A. Fleming, K. Geil, K. Jacobs, A. Jones, K. Knowlton, J. Koh, M. C. Lemos, J. Melillo, R. Pandya, T. C. Richmond, L. Scarlett, J. Snyder, M. Stults, A. M. Waple, J. Whitehead, D. Zarrilli, B. Ayyub M., J. Fox, A. Ganguly, L. Joppa, S. Julius, P. Kirshen, R. Kreutter, A. McGovern, R. Meyer, J. Neumann, W. Solecki, J. Smith, P. Tissot, G. Yohe, and R. Zimmerman, 2019: Evaluating knowledge to support climate action: A framework for sustained assessment. Report of an independent advisory committee on applied climate assessment. *Weather, Climate, and Society,* **11** (3), 465–487, https://doi.org/10.1175/WCAS-D-18-0134.1.
56. Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, **116** (12), 2417–2424, https://doi.org/10.1175/1520-0493(1988)116%3C2417:SSBOTM%3E2.0.CO;2.
57. Murphy, A. H., B. G. Brown, and Y. S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4** (4), 485–501, https://doi.org/10.1175/1520-0434(1989)004%3C0485:DVOTF%3E2.0.CO;2.
58. Murphy, A. H., 1991: Probabilities, odds, and forecasts of rare events. *Weather and forecasting*, **6** (2), 302–307, https://doi.org/10.1175/1520-0434(1991)006%3C0302:POAFOR%3E2.0.CO;2.
59. Murphy, A. H., 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather Forecasting*, **8** (2), 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.
60. Nearing, G. S. and H. V. Gupta, 2015: The quantity and quality of information in hydrologic models. *Water Resources Research*, **51** (1), 524–538, https://doi.org/10.1002/2014WR015895.
61. O'Neill, A. C., L. H. Erikson, and P. L. Barnard, 2017: Downscaling wind and wavefields for 21st century coastal flood hazard projections in a region of complex terrain. *Earth and Space Science*, **4** (5), 314–334, https://doi.org/10.1002/2016EA000193.

62. Pool, S., M. Vis, and J. Seibert, 2018: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, **63** (13–14), 1941–1953, https://doi.org/10.1080/02626667.2018.1552002.
63. RD 52.27.284-91. *Metodicheskie ukazaniya. Provedenie proizvodstvennyh (operativnyh) ispytanij novyh i usovershenstvovannyh metodov gidrometeorologicheskih i geliofizicheskih prognozov (RD 52.27.284-91. Methodical instructions. Carrying out production (operational) testing of new and improved methods of hydrometeorological and heliophysical forecasts).* Leningrad, Gidrometeoizdat, 150 p.
64. Reed, K. A., N. Goldenson, R. Grotjahn, W. J. Gutowski, K. Jagannathan, A. D. Jones, L. R. Leung, S. A. McGinnis, S. C. Pryor, A. K. Srivastava, P. A. Ullrich, and C. M. Zarzycki, 2022: Metrics as tools for bridging climate science and applications. *Wiley Interdisciplinary Reviews: Climate Change*, **13** (6), P. e799, https://doi.org/10.1002/wcc.799.
65. Roberts, N. M., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, **15** (1), 163–169, https://doi.org/10.1002/met.57.
66. Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Monthly Weather Review, **136** (1), 78–97, https://doi.org/10.1175/2007MWR2123.1.
67. Robertson, C., J. A. Long, F. S. Nathoo, T. A. Nelson, and C. C. Plouffe, 2014: Assessing quality of spatial models using the structural similarity index and posterior predictive checks. *Geographical Analysis*, **46** (1), 53–74, https://doi.org/10.1111/gean.12028.
68. Savage, N. H., P. Agnew, L. S. Davis, C. Ordóñez, R. Thorpe, C. E. Johnson, F. M. O'Connor, and M. Dalvi, 2013: Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation. *Geoscientific Model Development*, **6** (2), 353–372, https://doi.org/10.5194/gmd-6-353-2013.
69. Schoof, J. T., 2013: Statistical downscaling in climatology. *Geography Compass*, **7** (4), 249–265, https://doi.org/10.1111/gec3.12036.
70. Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, **379** (2194), https://doi.org/10.1098/rsta.2020.0097.
71. Shahriar, S. A., M. A. M. Siddique, and S. M. A. Rahman, 2021: Climate change projection using statistical downscaling model over Chittagong Division, Bangladesh. *Meteorology and Atmospheric Physics*, **133** (4), 1409–1427, https://doi.org/10.1007/s00703-021-00817-x.
72. Skok, G., 2015: Analysis of fraction skill score properties for a displaced rainband in a rectangular domain. *Meteorological Applications*, **22** (3), 477–484, https://doi.org/10.1002/met.1478.
73. Skok, G. and N. Roberts, 2018: Estimating the displacement in precipitation forecasts using the Fractions Skill Score. *Quarterly Journal of the Royal Meteorological Society*, **144** (711), 414–425, https://doi.org/10.1002/qj.3212.
74. Song, L., C. Xu, Y. Long, X. Lei, N. Suo, and L. Cao, 2022: Performance of seven gridded precipitation products over arid central Asia and subregions. *Remote Sensing,* **14** (23), 6039, https://doi.org/10.3390/rs14236039.
75. Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling,* **15** (1), 41–50, https://doi.org/10.1002/met.53.
76. Sun, Y., K. Deng, K. Ren, J. Liu, C. Deng, and Y. Jin, 2024: Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing,* **208**, 14–38, https://doi.org/10.1016/j.isprsjprs.2023.12.011.

77. Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres,* **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.
78. Tian, Y., G. S. Nearing, C. D. Peters-Lidard, K. W. Harrison, and L. Tang, 2016: Performance metrics, error modeling, and uncertainty quantification. *Monthly Weather Review*, **144** (2), 607–613, https://doi.org/10.1175/MWR-D-15-0087.1.
79. Vaughan, A., W. Tebbutt, J. S. Hosking, and R. E. Turner, 2022: Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, 15 (1), 251–268, https://doi.org/10.5194/gmd-15-251-2022.
80. Vrugt, J. A. and M. Sadegh, 2013: Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, **49** (7), 4335–4345, https://doi.org/10.1002/wrcr.20354.
81. Wang Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, 2004: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, **13** (4), 600–612, https://doi.org/10.1109/TIP.2003.819861.
82. Wehner, M. F., 2013: Very extreme seasonal precipitation in the NARCCAP ensemble: Model performance and projections. *Climate Dynamics*, **40**, 59–80, https://doi.org/10.1007/s00382-012-1393-1.
83. Weijs, S. V., G. Schoups, and N. Van De Giesen, 2010: Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, **14** (12), 2545–2558, https://doi.org/10.5194/hess-14-2545-2010.
84. Weniger, M., F. Kapp, and P. Friederichs, 2017: Spatial verification using wavelet transforms: a review. *Quarterly Journal of the Royal Meteorological Societ,* **143** (702), 120–136, https://doi.org/10.1002/qj.2881.
85. Wilby, R. L. and T. M. L. Wigley, 1997: Downscaling general circulation model output: a review of methods and limitations. *Progress in physical geography*, **21** (4), 530–548, https://doi.org/10.1177/030913339702100403.
86. Wilks, D. S., 2011: Statistical methods in the atmospheric sciences. Academic press, 676 p.
87. Willmott, C. J. and K. Matsuura, 2005: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, **30** (1), 79–82, https://doi.org/10.3354/cr030079.
88. Willmott, C. J., K. Matsuura, and S. M. Robeson, 2009: Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, **43** (3), 749–752, https://doi.org/10.1016/j.atmosenv.2008.10.005.
89. Yeggina, S., R. S. V. Teegavarapu, and S. Muddu, 2020: Evaluation and bias corrections of gridded precipitation data for hydrologic modelling support in Kabini River basin, India. *Theoretical and Applied Climatology*, **140**, 1495–1513, https://doi.org/10.1007/s00704-020-03175-7.
90. Zandler, H., T. Senftl, and K. A. Vanselow, 2020: Reanalysis datasets outperform other gridded climate products in vegetation change analysis in peripheral conservation areas of Central Asia. *Scientific Reports*, **10** (1), https://doi.org/10.1038/s41598-020-79480-y.
91. Zhong, X., F. Du, L. Chen, Z. Wang, and H. Li, 2024: Investigating transformer-based models for spatial downscaling and correcting biases of near-surface temperature and wind-speed forecasts. *Quarterly Journal of the Royal Meteorological Society*, **150** (758), 275–289, https://doi.org/10.1002/qj.4596.

Submitted 03.06.2024, accepted 10.12.2024.
For citation: Rezvov, V. Yu., M. A. Krinitskiy, and N. D. Tilinina, 2024: Point-by-point and complex quality metrics in atmosphere and ocean research: review of methods and approaches. *Journal of Oceanological Research*, **52** (4), 193–223, https://doi.org/10.29006/1564-2291.JOR-2024.52(4).10.