



Quantitative Auswertung der Sihl-Vorhersagesysteme für die Periode 2010-2025

GEO 511 Master's Thesis

Author: Arianna Manente, 17-712-076

Supervised by: Dr. Massimiliano Zappa (massimiliano.zappa@wsl.ch)

Faculty representative: Prof. Dr. Jan Seibert

24.10.2025

Abstract

Reliable hydrological forecasting is essential for flood risk management in urban areas with high exposure, such as the city of Zurich, where the Sihl River poses significant flood hazards. This study evaluates the predictive performance of nine MeteoSwiss weather forecast models for the Sihl catchment during the period 2010-2025, focusing on discharge thresholds relevant to flood management in Zurich. Across all model generations, forecast skill is highest at short lead times (24-48 h) and decreases significantly beyond 72 h. The Cosmo models, particularly Cosmo-2E, maintain positive probabilistic skill (BSS) up to the 40 m³/s threshold over five days, while the newer Icon models lose this advantage after 18 h. Persistent findings include a systematic underestimation of discharge, reduced discrimination at longer lead times, and calibration deficits and under-dispersion in the Icon models. Operationally, Cosmo-2E provided more reliable, decision-useful information for preventive actions, whereas Icon-2 offers value mainly in the short term. Improving climatological references, ensemble calibration, and cost-loss evaluations is essential to enhance the operational and economic utility of probabilistic flood forecasts in the Sihl system.

Danksagung

An dieser Stelle möchte ich mich herzlich bei allen bedanken, die zum Gelingen dieser Masterarbeit beigetragen haben. Mein besonderer Dank gilt Dr. Massimiliano Zappa von der Eidgenössischen Forschungsanstalt WSL für die engagierte Betreuung, die fachliche Unterstützung und die vielen wertvollen Anregungen während der gesamten Arbeit. Seine konstruktiven Rückmeldungen und seine stetige Hilfsbereitschaft haben massgeblich zum Fortschritt dieser Arbeit beigetragen. Ebenso danke ich Prof. Dr. Jan Seibert von der Universität Zürich (UZH) für die Betreuung seitens der Universität und das Interesse an meinem Forschungsvorhaben. Darüber hinaus möchte ich mich bei der Forschungsgruppe Gebirgshydrologie und Massenbewegungen an der WSL für die anregenden Gespräche, inspirierenden Diskussionen und die stets freundliche und offene Atmosphäre bedanken, die meine Zeit dort besonders bereichert hat. Ein besonderer Dank gilt auch meinen Freunden und meiner Familie für ihre Ermutigung und Unterstützung während dieser intensiven Zeit.

Inhalt

Abstract.....	1
Danksagung	2
1. Einleitung	5
1.1 Historische Entwicklung	5
1.2 Motivation.....	6
1.3 Operativer Kontext und Steuerung: IFKIS-Hydro Sihl.....	7
1.4 Praxisbeispiel: 40 m ³ /s-Schwelle am Limmatwehr	8
1.5 Zielsetzung und Forschungsfragen.....	9
1.6 Aufbau der Arbeit.....	9
2. Studiengbiet, Daten und Vorhersagesysteme	10
2.1 Gebiet und Messstellen.....	10
2.2 Beobachtungs- und Referenzdaten	11
2.3 Vorhersagemodelle	11
2.3.1 Umstellung von Cosmo zu Icon.....	12
3. Methodik.....	13
3.1 Datenbasis und Datenfilterung.....	13
3.2 Schwellendefinitionen.....	14
3.3 Metriken	15
3.3.1 Deterministische, kontinuierliche Metriken	16
3.3.2 Kontingenztafelbasierte Metriken	16
3.3.3 Probabilistische Metriken für binäre Ereignisse	17
3.3.4 Probabilistische Metriken für kontinuierliche Grössen	17
4. Ergebnisse	18
4.1 Überblick und Zielsetzung.....	18
4.2 Deterministische, kontinuierliche Metriken	19
4.2.1 Allgemeine Modellgüte	19
4.2.2 Bias und systematische Fehler.....	23
4.2.3 Lineare Zusammenhänge.....	23
4.3 Ereignisbezogene Verifikation und Schwellenanalyse.....	26
4.3.1 Allgemeine Ereignisstatistik	26
4.4 Probabilistische Metriken für binäre Ereignisse	29
4.4.1 Gesamtgüte der probabilistischen Vorhersagen	29
4.4.2 Diskriminationsfähigkeit der meteorologischen Modelle	31
4.4.3 Zuverlässigkeit der Wahrscheinlichkeiten	34

4.5	Probabilistische Metriken für kontinuierliche Grössen	36
4.5.1	Gesamtgüte der Ensembleverteilung	36
4.5.2	Zeitliche Entwicklung der probabilistischen Modellgüte	38
4.5.3	Probabilistischer Skill relativ zur Klimatologie.....	40
4.5.4	Kalibrierung der Ensembleverteilungen.....	42
4.6	Auswertungen für den Sihlsee.....	43
5.	Diskussion.....	45
5.1	Gesamtbewertung der Modellgüte.....	45
5.2	Ereignisbasierte Vorhersage und Schwellenanalyse	46
5.3	Probabilistische Bewertung	47
6.	Ökonomische Bewertung und Entscheidungsrelevanz.....	50
7.	Schlussfolgerungen und Ausblick	51
8.	Literaturverzeichnis	53
9.	Anhang.....	56
9.1	Formeln und Definitionen der verwendeten Gütemassen	56
9.1.1	Deterministische, kontinuierliche Gütemasse	56
9.1.2	Kontingenztafelbasierte Metriken	57
9.1.3	Probabilistische Metriken für binäre Ereignisse	59
9.1.4	Probabilistische Metriken für kontinuierliche Grössen	61
9.1.5	Ensembleverteilung	62
9.2	Reproduzierbarkeit.....	62
9.3	Erklärung der Nutzung von Künstlicher Intelligenz.....	62
9.4	Zusatzabbildungen	63
9.5	Eidesstattliche Erklärung	70

1. Einleitung

1.1 Historische Entwicklung

Historisch wird die Bedeutung hydrologischer Vorhersagesysteme für Katastrophenschutz und Wasserressourcenmanagement bereits 1975 von der World Meteorological Organization (WMO) hervorgehoben (WMO, 1975). In den vergangenen vier Jahrzehnten entwickelte sich die Abflussvorhersage von deterministischen Einzellösungen zu ensemblebasierten Verfahren (Demeritt et al., 2007; Troin et al., 2021). Die Literatur fasst die operationelle Entwicklungslinie der Ensemble-Abflussvorhersagen in drei Hauptklassen: (1) statistikbasierte Systeme, die auf historischen Abflussdaten beruhen, (2) Ensemble Streamflow Prediction (ESP), die mit historischen meteorologischen Szenarien arbeiten und (3) Ensemble Prediction Systems (EPS), bei denen numerische Wettervorhersagen (Numerical Weather Prediction NWP) mit hydrologischen Modellen gekoppelt werden (Demeritt et al., 2007; Troin et al., 2021). Diese probabilistischen Verfahren wurden zunächst im European Centre for Medium-Range Weather Forecasting (ECMWF) entwickelt und ab Mitte der 1990er Jahre operationell eingesetzt (Demeritt et al., 2007). Ab den späten 1990er und frühen 2000er Jahren verlagerte sich der Fokus zunehmend auf kürzere Vorhersagehorizonte, die direkte Kopplung von NWP und Hydrologiemodelle sowie die explizite Quantifizierung von Prognoseunsicherheit (Demeritt et al., 2007; Troin et al., 2021). Die Prognoseunsicherheit wird quantifiziert, indem eine Bandbreite (Ensemble) möglicher zukünftiger Abflussprognosen generiert werden (Demargne et al., 2014). Parallel zum methodischen Ausbau rückte die operationelle Nutzung (unter anderem für Warnung und Einsatzplanung) stärker in den Fokus. Initiativen wie «Hydrological Ensemble Prediction Experiment» (HEPEX, seit 2004) fördern den Austausch zwischen Forschung und Vorhersagediensten und unterstützen die Aufbereitung probabilistischer Produkte für Entscheidungen (Thielen et al., 2008; Troin et al., 2021). Zusammengefasst ist Ensemble-Hydrologie heute ein etabliertes Feld, dessen Mehrwert insbesondere in der Quantifizierung der Gesamtunsicherheit gesehen wird (Rossa et al., 2011). Wie Zappa et al. (2011) zeigen, ermöglicht die Kombination von Ensemble-Wettervorhersagen, ensemblebasierten Beobachtungsdaten und hydrologischen Modellunsicherheiten eine realistische Quantifizierung der gesamten Prognoseunsicherheit in Hochwasservorhersagen.

In der Schweiz nahm die digitale hydrologische Modellierung in den 1970er Jahren ihren Anfang (Horton et al., 2022). Der Schwerpunkt lag zunächst auf der Simulation von Abflüssen infolge von Schnee- und Gletscherschmelze (Braun, 1984; Braun & Aellen, 1990; Martinec, 1970), wodurch die Schneehydrologie die Modellentwicklung über mehrere Jahrzehnte prägte (Horton et al., 2022). Die quantitativen Echtzeitprognosen für die Wasserkraftproduktion und Wasserressourcenmanagement lösten in dieser Zeit statistische Verfahren zunehmend durch prozessbasierte hydrologische Modelle ab (Horton et al., 2022; Lugiez et al., 1969). Frühe Anwendungen und fallbezogene Modellanpassungen, wie die HBV-Weiterentwicklungen für Schweizer Einzugsgebiete durch Hottel et al. (1992) verdeutlichen diesen Übergang. Ab den 1980er und 1990er Jahren wurden in der Schweiz erste operationelle Hochwasservorhersagen eingeführt, die auf deterministischen Modellen beruhten und zunächst vorwiegend der Schifffahrt dienten (Zappa & Vogt, 2007). Mit dem Fortschritt in der numerischen Wettervorhersage und der Radartechnologie entstanden ab den 2000er-Jahren zunehmend datengetriebene, automatisierte Prognosesysteme (Zappa & Vogt, 2007). Einen wichtigen Entwicklungsschritt stellte das von MeteoSchweiz koordinierte Projekt MAP D-PHASE dar, in dem erstmals atmosphärische Ensemble-Vorhersagen mit hydrologischen Modellen gekoppelt wurden, um Unsicherheiten in alpinen Hochwasserprognosen systematisch zu berücksichtigen (Zappa et al., 2008). Dieses internationale Demonstrationsprojekt markierte den Übergang von deterministischen zu probabilistischen Verfahren in der operationellen Vorhersagepraxis (Zappa et al., 2008). Ein

prominentes Beispiel für die frühe operative Kopplung atmosphärischer und hydrologischer Ensembles ist das seit 2008 betriebene System für die Sihl (Addor et al., 2011), in dem probabilistische und deterministische Wettervorhersagen mit einer hydrologisch-hydraulischen Kette verknüpft werden (siehe Kapitel 1.3).

1.2 Motivation

Die Stadt Zürich liegt auf dem Schwemmkegel der Sihl, einem natürlichen Überschwemmungsgebiet, und ist durch extreme Hochwasserereignisse stark gefährdet (Abbildung 1). Bei Extremereignissen drohen grossflächige Überflutungen in der Innenstadt und des Hauptbahnhofs (HB). Das Schadenspotenzial wird auf über 6.7 Mrd. CHF geschätzt, ohne volkswirtschaftliche Folgeschäden wie Betriebsstörungen, Unterbrüche und die Zerstörung der Infrastruktur für Energie, Telekommunikation und Verkehr zu berücksichtigen (AWEL, 2017). Besonders kritisch ist die Engstelle am HB, wo die Sihl durch fünf parallele Schächte zwischen dem Tiefbahnhof und den oberirdischen Gleisen geführt wird. Die maximale Durchflusskapazität dieser Passage beträgt $350 \text{ m}^3/\text{s}$, was einem 100-jährigen Hochwasser entspricht (Liechti et al., 2016). Dadurch stellt der Hauptbahnhof eine hydraulische Engstelle im Abflusssystem dar.



Abbildung 1: Schwemmkegel der Sihl (AWEL, 2017).

1.3 Operativer Kontext und Steuerung: IFKIS-Hydro Sihl

Angesichts des potenziellen Schadensausmasses im Stadtgebiet ist für das urbane Risikomanagement ein verlässliches Hochwasservorhersagesystem von zentraler Bedeutung. Seit 2007 läuft IFKIS-Hydro Sihl, ein regionales Warnsystem zur Verbesserung der Hochwassersicherheit des Sihltals und der Stadt Zürich (Abbildung 2). Es verbindet deterministische und probabilistische Wettermodelle von MeteoSchweiz mit einer gekoppelten hydrologisch-hydraulischen Modellkette (PREVAH-FLORIS) (Addor et al., 2011; Zappa et al., 2010). Es wird von der Eidgenössischen Forschungsanstalt für Wald, Schnee und Landschaft (WSL) im Auftrag der Zürcher kantonalen Fachstelle, Amt für Abfall, Wasser, Energie und Luft (AWEL), betrieben. Ziel ist, den Informationsgewinn im Hinblick auf Hochwasserrisiken durch Ensembleprognosen zu nutzen und Entscheidungen von Behörden, Infrastrukturbetreibern und Notfallmanagement zu stützen.

Im operationellen Kontext umfasst der Hochwasserschutz des Kantons Zürich neben baulichen Vorkehrungen insbesondere die Vorabsenkung des Sihlsees, um bei prognostizierten kritischen Niederschlagsereignissen den Stausee als Retentionsraum zu nutzen und Abflussspitzen in der Stadt zu dämpfen. Eine Vorabsenkung wird in Betracht gezogen, wenn die Vorhersagen, unter Einbezug des aktuellen Seestands, ein Überschreiten des Stauziels 889.34 m ü. M. erwarten lassen oder eine Pegelzunahme von mehr als 2 cm pro 30 Minuten ab 888.7 m ü. M. wahrscheinlich ist (Badoux et al., 2010; Etzelwerk AG, 2017). Diese Situationen lösen gemäss Wehrreglement eine automatische Hochwasserentlastung des Stausees aus, wodurch die Wassermengen des Sihlsees ungünstig mit allfälligen Hochwasserspitzen aus den unteren Zuflüssen der Sihl, Alp und Biber, zusammenfallen können (Etzelwerk AG, 2017). Um dies zu vermeiden, wird der Sihlsee vorsorglich abgesenkt, sodass ausreichend Retentionsraum zur Verfügung steht und die Seeabgabe zeitlich entkoppelt erfolgen kann. Die Entlastung muss dafür ca. 24-48 Stunden vor dem Ereignisbeginn, in Extremfällen bis spätestens 72 Stunden vorher, erfolgen (Badoux et al., 2010). Weil das Timing und die Intensität der Vorabsenkung direkt von der Qualität der Vorhersagen abhängen, ist die Bewertung der Prognosegüte zentral für die operative Hochwassersicherheit. Die zugrunde liegende Datenbasis (Abfluss- und Seepegelmessung, Radar- und Stationsniederschläge) sowie die Vorhersagesysteme werden in Kapitel 2 beschrieben. Wie stark sich Prognosen unmittelbar auf operative Schwellenentscheide im Stadtgebiet auswirken, zeigt sich besonders an der Baustelle beim Limmatwehr.

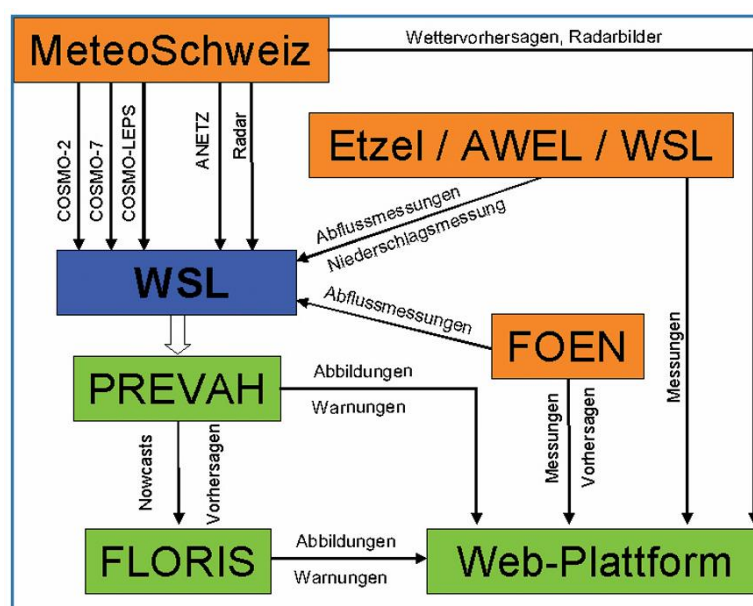


Abbildung 2: Schematische Darstellung der Vorhersagekette von IFKIS-Hydro Sihl. Cosmo-Modelle wurden mittlerweile durch Icon-Modelle ersetzt (Zappa et al., 2010).

1.4 Praxisbeispiel: 40 m³/s-Schwelle am Limmatwehr

Am Limmatwehr beim Zusammenfluss von Sihl und Limmat, wenige 100 m unterhalb des HB, laufen zwischen 2024 und 2028 Bauarbeiten (AWEL, 2019), die direkt von der Prognosequalität der Wettermodelle abhängen. Für die Bauarbeiten wurden Kiesbänke am Sihlufer entlang, sowie eine provisorische Brücke für die Zufahrt der Baufahrzeuge aufgebaut (Abbildung 3). Bereits bei 40 m³/s Abfluss der Sihl muss die provisorische Brücke gesperrt und Bauhilfsmittel entfernt werden, da die Durchfliesskapazität unter der Brücke erreicht ist. Ab einem Abfluss von 80 m³/s sind Stahlelemente zu demontieren und 160–180 m³/s Abfluss erzwingen den Rückbau ganzer Zufahrten. Diese Eingriffe sind aufwändig und kostenintensiv und erfordern mindestens einen Tag Vorlaufzeit. Dies bedeutet, dass basierend auf hydrologischen Vorhersagen um 8 Uhr morgens entschieden werden muss, ob die Baustelle bis zum Nachmittag gesperrt oder sogar geräumt werden muss. In der Vergangenheit führten Fehlalarme (Modelle hatten einen Abfluss von 40 m³/s oder höher prognostiziert, jedoch wurde dieser in der Realität nicht erreicht) zu einem Vertrauensverlust in die Modelle. Genau hier setzt diese Arbeit an, denn die Prognosegüte der hydrologischen Vorhersagemodelle des Einzugsgebiets der Sihl sind für den Hochwasserschutz operativ entscheidend.



Abbildung 3: Provisorische Brücke für Baufahrzeuge, sowie Kiesbänke für Zufahrt.

1.5 Zielsetzung und Forschungsfragen

Diese Arbeit bewertet die Prognosegüte der meteorologischen Vorhersagemodelle für das Einzugsgebiet der Sihl im Zeitraum 2010-2025. Dabei steht die Frage im Fokus, wie zuverlässig diese Modelle kritische Abflussschwellen vorhersagen können. Es genügt nicht, dass ein Wettermodell im Mittel richtig liegt, entscheidend ist, ob Extremwerte wie Hochwasserereignisse zuverlässig erkannt werden. Die Relevanz für den Hochwasserschutz liegt genau in diesen Extremen.

Im Rahmen dieser Arbeit werden vier zentrale Forschungsfragen behandelt:

- Wie hat sich die Vorhersagegüte der Modelle im Laufe der Jahre entwickelt und welche Faktoren beeinflussen die Präzision der Prognosen?
- Welche Fehlerquellen und Unsicherheiten bestehen innerhalb der Modelle, insbesondere im Hinblick auf statistische Parameter wie Korrelation und probabilistische Bewertungsskalen?
- Inwieweit lassen sich durch verbesserte hydrologische Vorhersagen wirtschaftliche Vorteile für den Hochwasserschutz quantifizieren?
- Wie zuverlässig können die Modelle spezifische, für Endnutzer relevante Schwellenwerte - wie etwa den kritischen Abfluss von 40 m³/s beim Limmatwehr in Zürich - vorhersagen und wie verhalten sich die Prognosegüte bei der Überschreitung solcher Schwellenwerte?

1.6 Aufbau der Arbeit

Der Aufbau der Arbeit folgt einer klaren inhaltlichen Struktur: Zunächst werden in Kapitel 2 das Untersuchungsgebiet, die Datengrundlage und die verwendeten Vorhersagemodelle vorgestellt. Anschliessend erläutert Kapitel 3 die methodischen Vorgehensweisen. Die daraus resultierenden Ergebnisse werden in Kapitel 4 präsentiert und in Kapitel 5 kritisch diskutiert. Kapitel 6 widmet sich der ökonomischen Bewertung, bevor Kapitel 7 die wichtigsten Erkenntnisse zusammenfasst. Detaillierte Ergänzungen finden sich in den Anhängen.

2. Studiengebiet, Daten und Vorhersagesysteme

2.1 Gebiet und Messstellen

Die Sihl entspringt südwestlich des Zürichsees in den Schwyzer Voralpen auf ca. 1'900 m ü. M. und mündet bei Zürich auf etwa 404 m ü. M. in die Limmat (Abbildung 4). Ihr Einzugsgebiet umfasst 336 km² und die mittlere Geländehöhe beträgt 1'060 m ü. M. Zur Stromgewinnung wurde 1937 der Sihlstausee in Betrieb genommen. Der Stausee prägt das System hydrologisch: Er teilt das Gesamteinzugsgebiet am Stausee-Auslass Schlagen in zwei hydrologische Hauptteile, das obere Einzugsgebiet (155 km²) im Süden und das untere nördliche Einzugsgebiet bis Zürich (181 km²). Seit der Inbetriebnahme des Stausees werden die Abflüsse der oberen Sihl, der Minster und kleinerer Zuflüsse bei etwa 889 m ü. M. im Sihlsee zwischengespeichert und zu einem grossen Teil zur Stromerzeugung in den Zürichsee geleitet. Durch den Sihlsee steht im oberen Einzugsgebiet ein grosser, aktiv steuerbarer Speicher zur Verfügung, der bei kritischer Lage zur temporären Dämpfung von Abflussspitzen genutzt werden kann. Die operative Umsetzung und Entscheidungskriterien sind in Kapitel 1.3 beschrieben. Alp und Biber fliessen unterhalb des Sihlsees in die Sihl und gehören zum unteren Einzugsgebiet. Etwa 1.4 km oberhalb des HB, auf 412 m ü. M., wird am Standort Zürich-Sihlhölzli der stündliche Abfluss der Sihl gemessen. Insgesamt stehen dem IFKIS-Hydro Sihl die Echtzeit-Abflussdaten von sechs Messstationen zur Verfügung. (Kleinn et al., 2019; Zappa et al., 2010)

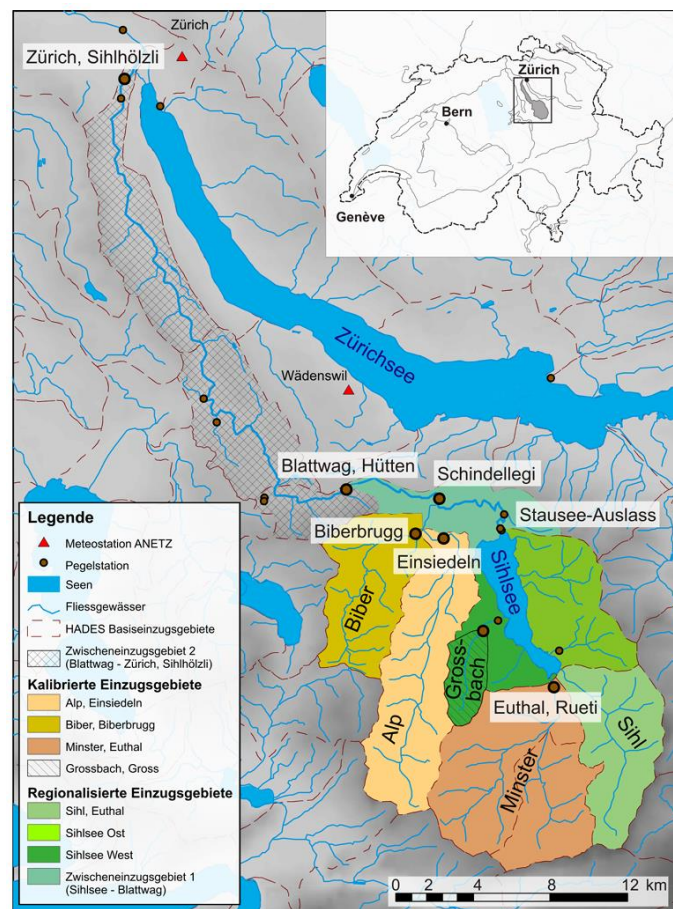


Abbildung 4: Übersichtskarte des Einzugsgebietes der Sihl (Schwanbeck et al., 2007).

2.2 Beobachtungs- und Referenzdaten

Für diese Arbeit werden die hydrologischen Beobachtungsdaten von zwei der offiziellen Pegelmessstationen des Bundesamts für Umwelt (BAFU) sowie der Kantonalen Fachstelle AWEL genutzt. Die Station Sihlhölzli, welche vom BAFU betrieben wird, befindet sich im Stadtgebiet und ist besonders relevant für die Beurteilung hochwasserrelevanten Risiken im städtischen Umfeld. Die Station beim Sihlsee misst den Seepiegel, welcher die Abflüsse flussabwärts beeinflusst und insbesondere für die Interpretation längerfristiger Vorhersagen relevant ist. Die Beobachtungsdaten des Seepiegels wurden vom AWEL und die Daten des Abflusses vom BAFU für den Zeitraum 2010 bis August 2025 zur Verfügung gestellt.

2.3 Vorhersagemodelle

Für die quantitative Analyse der Vorhersagegüte wurden neun meteorologische Modelle von MeteoSchweiz berücksichtigt, welche zwischen 2010 und 2025 operationell waren. Numerische Wettermodelle stützen sich auf physikalische Gesetze und mathematische Gleichungen, um zentrale physikalische Prozesse in der Atmosphäre und am Erdboden, wie Luftdruck, Temperatur, Wind, Feuchtigkeit, Wolkenbildung und Niederschlag, zu beschreiben (MeteoSchweiz, 2025b). Abbildung 5 bietet einen Überblick der neun meteorologischen Modelle mit den jeweiligen Eigenschaften. Sie unterscheiden sich in räumliche Auflösung (zwischen $1.1 \times 1.1 \text{ km}^2$ und $7 \times 7 \text{ km}^2$), Anzahl Initialisierungen (Beispiel 8 x täglich mit Start 00, 03, ... 21 UTC), Laufzeit (Vorhersage von 24 h bis 132 h), Anzahl Member (1-21), Einsatzzeitraum und Anzahl der WSL verfügbaren Vorhersagen. Wettermodelle, welche in der gleichen Periode aktiv waren, haben die gleiche Farbe. Die Ablösung der Vorgängermodelle durch neuere Modelle findet von links nach rechts statt. Cosmo-2 wurde 2016 von Cosmo-1 abgelöst, welcher wiederum 2020 von Cosmo-1E abgelöst wurde, bis im Sommer 2024 Icon-1 Cosmo-1E abgelöst hat. Es gibt deterministische und probabilistische Wettermodelle. Die deterministischen Wettermodelle sagen für jeden Zeitschritt einen einzigen Prognosewert vorher, während probabilistische Wettermodelle für jeden Zeitschritt mehrere gleich wahrscheinliche Prognosewerte vorhersagen (Cloke & Pappenberger, 2009; Liechti et al., 2016). Aus diesen gleich wahrscheinlichen Prognosewerten kann die Eintrittswahrscheinlichkeit für ein bestimmtes Ereignis (wie das Überschreiten von $40 \text{ m}^3/\text{s}$ beim Limmatwehr) berechnet werden (MeteoSchweiz, 2025a). Die Streuung der unterschiedlichen Prognosewerte eines Ensemblemodells ermöglicht es, die Unsicherheiten der Vorhersagen der Wetterlage besser abzuschätzen (MeteoSchweiz, 2025b). Unter Laufzeit versteht man die Zeitspanne, die ein Modell ab dem Startzeitpunkt in die Zukunft abdeckt und Initialisierung beschreibt die Anzahl täglicher Durchrechnungen der Modelle. Icon-1 berechnet achtmal täglich (Initialisierung), also alle drei Stunden, die neue Entwicklung des Wetters für eine Vorhersagezeit (Vorlaufzeit) von bis zu 32 Stunden (MeteoSchweiz, 2025a). Die 11 Member von Icon-1 rechnen pro Zeitschritt mit leicht unterschiedlichen Anfangsbedingungen 11 gleich wahrscheinliche, unabhängige Prognosewerte (MeteoSchweiz, 2025a). Somit entstehen pro Tag 11 Prognosen im drei-Stunden-Takt für die nächsten 32 Stunden durch Icon-1. Je kleiner die Streuung der verschiedenen Prognosen ist, desto sicherer ist die Vorhersage des meteorologischen Modells.

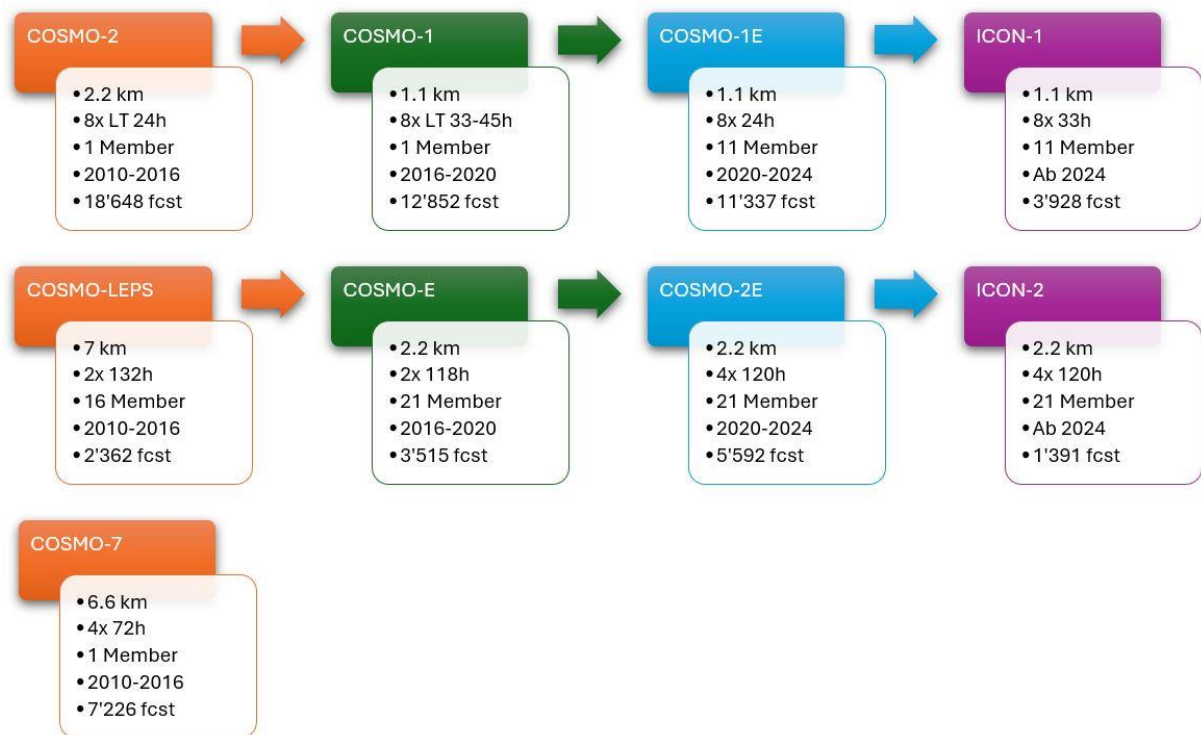


Abbildung 5: Schematische Darstellung der MeteoSchweiz Wettermodelle mit Namen, räumliche Auflösung, Anzahl Initialisierungen, Laufzeit, Anzahl Member, Einsatzzeitraum und Anzahl verfügbare Vorhersagen. Die Ablösung durch neuere Modelle findet von links nach rechts statt.

2.3.1 Umstellung von Cosmo zu Icon

Zwischen 2010 und 2024 waren verschiedene Generationen von Cosmo operationell im Einsatz, bis MeteoSchweiz 2024 das Prognosemodell Icon operativ eingeführt hat. Folgend werden die Unterschiede zwischen Icon und Cosmo aufgeführt. Während Cosmo auf ein viereckiges Gitter basiert, modelliert Icon die Schweiz mit einem dreieckigen Gitter (Abbildung 6). Dies führt zu einer genaueren Modellierung, was insbesondere bei der komplexen Topografie der Alpen hilfreich ist. Icon ermittelt die für die Prognosen entscheidenden Prozesse und den künftigen Atmosphärenzustand mit verbesserten numerischen Methoden (MeteoSchweiz, 2025b).

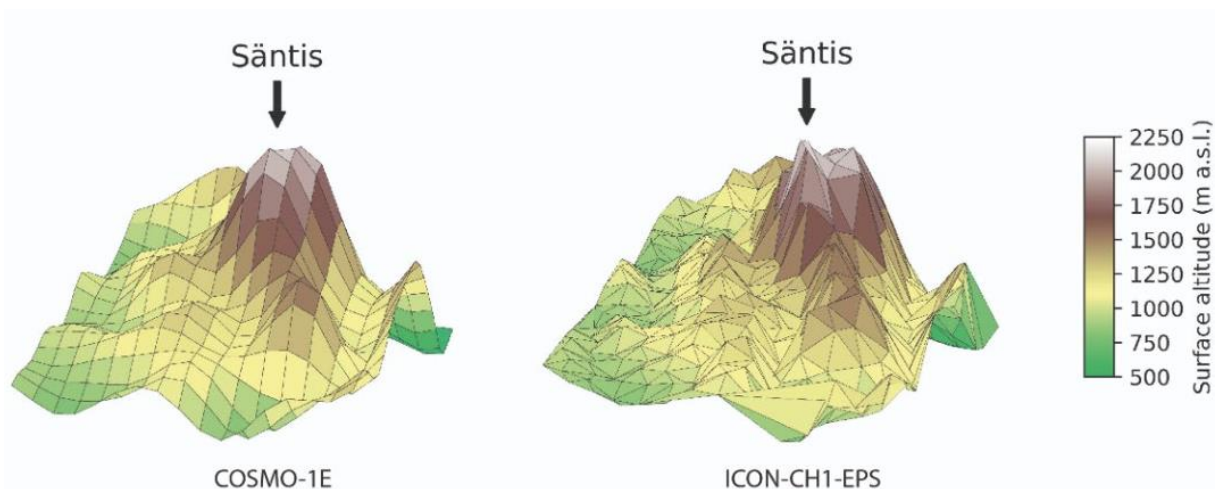


Abbildung 6: Unterschiedliche Gitterstruktur der meteorologischen Modelle Cosmo und Icon (MeteoSchweiz, 2025b).

3. Methodik

Die quantitative Bewertung der neun Vorhersagesysteme für die Sihl erfolgt auf einer konsistenten Auswertebasis für die hochwasserrelevante Saison und einem Set aus deterministischen sowie probabilistischen Gütemassen. Die Auswahl der Metriken folgt der in der Hydrologie etablierten Verifikationspraxis wie von Brown et al. (2010) beschrieben. Grundlegend für Definitionen und Interpretation sind Wilks' (2006) Darstellungen zu Verifikationsmassen und Kalibrierung und Diskrimination.

3.1 Datenbasis und Datenfilterung

Analysiert wird die Saison Mai bis November aller Jahre in der Studienperiode. Diese Einschränkung entspricht dem hydrologischen Risikofenster für lang anhaltende Niederschläge, Gewittern sowie Schmelzereignisse im Frühsommer (Liechti et al., 2016). Während 21. - 28. Juli 2014 sowie 8. - 15. Juli 2021 wurde der Sihlsee präventiv vor einem starken Gewitter gesenkt (siehe Hochwasserschutzmassnahmen in Kapitel 1.3). Zur Vermeidung anthropogener Verzerrungen der Modellprognosen werden diese zwei Eingriffe konsequent aus allen Auswertungen ausgeschlossen.

Die Datenverfügbarkeit wird pro meteorologischem Modell und Vorlaufszeit über eine Ausfallquote (Anteil fehlender Werte) kontrolliert. Als Kriterium gilt maximal 30 % fehlende Werte, was einem ≥ 70 % Verfügbarkeit entspricht. Fälle, die das Kriterium verfehlen, werden von der jeweiligen Auswertung ausgeschlossen. Für den Sihlseepegel wird in allen Metriken eine Mess-, beziehungsweise Betriebstoleranz von ± 5 cm berücksichtigt. Abweichungen innerhalb dieser Bandbreite werden als messtechnisch unbedeutend gewertet und von der Fehlerzählung ausgeschlossen. Die Toleranz wurde in allen Metriken konsistent angewandt. Für einige Metriken werden die Vorlaufszeiten in fünf disjunkte 24-Stunden-Bins zusammengefasst: Tag 1 = 1-24 h, Tag 2 = 25-48 h, Tag 3 = 49-72 h, Tag 4 = 73-96 h, Tag 5 = 97-120 h. Der operationelle Kontext (PREVAH–FLORIS-Kette in Abbildung 2, die Rolle des Sihlsees und typische Vorwarnzeiten) wird durch die IFKIS-Berichte (Badoux et al., 2010; Zappa et al., 2010) und Sihl-Studien (Addor et al., 2011; Liechti et al., 2016; Liechti & Zappa, 2012, 2014, 2017, 2020) abgedeckt und motivieren diese Einstellungen und die Fokussierung auf schwellenbasierte Verifikation aus Endnutzerperspektive.

3.2 Schwellendefinitionen

Ein Ereignis wird definiert, wenn ein beobachteter oder vorhergesagter Wert die jeweilige Schwelle überschreitet (Wert $\geq$ Schwelle). Für die stationäre Schwellenverifikation werden bei der Abflussmessstation Sihlhölzli und der Pegelmessstation beim Sihlsee die in Tabelle 1 zusammengefasste Schwellen verwendet. Die besonders für Endnutzer relevante Abflussschwelle 40 m³/s am Limmatwehr entspricht der operativen Eingriffsschwelle (siehe Kapitel 1.4). Die Schwellen des Sihlsees basieren auf die Betriebsregeln der Etzelwerk AG (2017): bei Normalbetrieb wird ein Stauziel von 889.34 m. ü. M. für die Energieerzeugung angestrebt. Wenn der Seepiegel über das Stauziel hinaus oder mehr als 2 cm in 30 Minuten über 888.7 m. ü. M. ansteigt, tritt eine manuelle oder automatische Seeregulierung in Kraft. 889 m ü. M. ist ein Zwischenwert. Ein Seepiegel von 887.34 m ü. M. beschreibt die Mückengrenze, 888.34 m ü. M. liegt 1m darüber. Zur Vergleichbarkeit zwischen Vorlaufzeiten wurde aus den Beobachtungsdaten zunächst eine empirische Verteilungsfunktion gebildet. Damit wurden die physikalischen Schwellen in Wahrscheinlichkeiten übersetzt und anschliessend pro Vorlaufzeit die entsprechenden Beobachtungsquantile bestimmt.

Tabelle 1: Übersicht der für die Verifikation definierten Schwellen.

Schwellen					
Sihlhölzli	10 m ³ /s	20 m ³ /s	40 m ³ /s	60 m ³ /s	80 m ³ /s
Sihlsee	887.34 m ü. M.	888.34 m ü. M.	888.70 m ü. M.	889.00 m ü. M.	889.34 m ü. M.

3.3 Metriken

Tabelle 2 gibt eine Orientierung über alle verwendeten Metriken, während im Anhang 9.1 die vollständigen Definitionen und Formeln nach Wilks (2006) aufgeführt sind. Die Verifikation erfolgt sowohl mit kontinuierlichen als auch mit schwellenbasierten Gütemassen.

Tabelle 2: Übersichtstabelle aller verwendeten Metriken mit Name und kurzer Definition nach Wilks (2006). Vollständige Formeln und Quellen siehe Kapitel 9.1.

	Metrik	Kurzdefinition
Deterministische Metriken	Kontinuierliche, deterministische Metriken	
	Mean Error (ME)	Differenz zwischen der durchschnittlichen Vorhersage und der durchschnittlichen Beobachtung
	Mean Absolute Error (MAE)	Mittlere absolute Abweichung
	Root Mean Square Error (RMSE)	Abweichung zwischen Modellvorhersagen und tatsächlichen Beobachtungen
	Pearson's correlation coefficient (r)	Lineare Beziehung zwischen Vorhersage und Beobachtung
	Coefficient of determination (R^2)	Anteil der Varianz, erklärt durch den linearen Zusammenhang
	Nash-Sutcliffe Efficiency (NSE)	Vergleich Vorhersagevarianz mit Beobachtungsvarianz
	Kontingenztafel-basierte Metriken	
	Probability of Detection (POD)	Anteil erkannter Ereignisse
	False Alarm Ratio (FAR)	Anteil Fehlalarme an allen Alarmen
	Probability of False Detection (POFD)	Anteil falscher Alarme an allen Nicht-Ereignissen
Probabilistische Metriken	Critical Success Index (CSI)	Trefferquote ohne korrekte Neins
	Heidke Skill Score (HSS)	Vorhersagevorteil gegenüber Zufall
	Probabilistische Metriken für binäre Ereignisse	
	Brier Score (BS)	Mittlerer quadratischer Fehler zwischen Vorhersage und Beobachtung
	Brier Skill Score (BSS)	Skill des BS relativ zur Referenz (Klimatologie)
	Receiver Operating Characteristic (ROC)	Visualisiert die Diskriminationsfähigkeit eines Modells über alle Entscheidungsschwellen
	Area Under the Curve (AUC)	Quantifiziert die Diskriminationsfähigkeit, die im ROC-Diagramm dargestellt wird
	Reliability-Diagramm	Kalibrierung: beobachtete vs. vorhergesagte Wahrscheinlichkeit
	Attribute Diagram	Reliability Diagram mit zusätzlichen Referenzlinien
	Probabilistische Metriken für kontinuierliche Größen	
	Continuous Ranked Probability Score (CRPS)	Distanz zwischen vorhergesagter und beobachteter Verteilungsfunktion
	Continuous Ranked Probability Skill Score (CRPSS)	Skill des CRPS relativ zur Referenz (Klimatologie)
	Ensemble Diagnostik	
	Rang-Histogramm	Kalibrierung beziehungsweise Konsistenz von Ensemblevorhersagen

3.3.1 Deterministische, kontinuierliche Metriken

Zur quantitativen Bewertung der deterministischen Modellgüte wurden für jede Station, jedes Jahr und jede Vorlaufzeit vier klassische kontinuierliche Gütemasse berechnet: Mean Error (ME), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) und die Nash-Sutcliffe-Effizienz (NSE) (vgl. Kapitel 9.1.1.1-9.1.1.6). Bei Ensemblemodellen wurde für alle deterministischen Metriken der Median der Ensemblemitglieder als repräsentativer Vorhersagewert verwendet, um eine robuste Vergleichsbasis mit den rein deterministischen Wettermodellen zu schaffen. Alle Metriken wurden pro Vorlaufszeit und Jahr gespeichert, um sowohl die Abnahme der Vorhersagegüte mit wachsender Vorlaufzeit als auch längerfristige Trends zwischen den Modellgenerationen darstellen zu können. Die Berechnung der Metriken erfolgte nach den Standarddefinitionen (vgl. Wilks, 2006):

- Der ME quantifiziert die mittlere Abweichung und zeigt systematische Über- oder Unterschätzungen.
- Der MAE misst die mittlere absolute Abweichung und ist robust gegenüber Ausreißern.
- Der RMSE betont grössere Fehler stärker und dient als Mass für die Gesamtstreuung der Prognosefehler.
- Die NSE vergleicht die Modellvarianz mit der Beobachtungsvarianz und bewertet den Reproduktionsgrad der beobachteten Dynamik.

Zur Beurteilung der linearen Übereinstimmung zwischen Beobachtungen und Vorhersagen wurde ergänzend das Quadrat des Pearson-Korrelationskoeffizienten (Bestimmtheitsmass R^2) berechnet. Der Wert beschreibt den Anteil der beobachteten Varianz, der durch die Modellvorhersage erklärt wird, und liegt zwischen 0 (kein Zusammenhang) und 1 (perfekte lineare Übereinstimmung) (Wilks, 2006). Für jede Vorlaufszeit wurde der R^2 -Wert zwischen beobachteter und vorhergesagter Zeitreihe bestimmt.

3.3.2 Kontingenztafelbasierte Metriken

Zur Bewertung der Ereignisvorhersage wurden schwellenbasierte Kontingenzstatistiken berechnet. Ein Ereignis ist definiert, wenn der beobachtete oder vorhergesagte Wert eine festgelegte Schwelle überschreitet (vgl. Abschnitt 3.2). Für jedes meteorologische Modell, jede Vorlaufszeit und jeden Schwellenwert wurde eine 4-Felder-Kontingenztafel erstellt (vgl. Abschnitt 9.1.2). Aus diesen Zählungen wurden eine Reihe standardisierter Scores abgeleitet (vgl. Wilks, 2006):

- Critical Success Index (CSI) als Mass der Trefferquote ohne Berücksichtigung korrekter Neins,
- Heidke Skill Score (HSS) zur Bewertung des Skillvorteils gegenüber einem Zufallsmodell,
- False Alarm Ratio (FAR) als Anteil der Fehllalarme an allen Alarmausgaben,
- Probability of Detection (POD) und Probability of False Detection (POFD) zur Charakterisierung von Sensitivität und Fehllarmneigung

Um die Ereignisvorhersage der unterschiedlichen Wettermodelle zu vergleichen, wurde eine Aufteilung und Gruppierung der Modelle nach drei Kategorien vorgenommen. Einerseits wurden die meteorologischen Modelle je nach hoher ($1.1 \times 1.1 \text{ km}^2$) oder tiefer ($> 2.1 \times 2.2 \text{ km}^2$) räumlicher Auflösung aufgeteilt. Andererseits wurden die meteorologischen Modelle nach kurzer ($< 33 \text{ h}$) oder langer ($> 34 \text{ h}$) Vorlaufszeit aufgelöst. Ausserdem wurde je eine Gruppe für deterministische und probabilistische Wettermodelle gebildet.

3.3.3 Probabilistische Metriken für binäre Ereignisse

Zur Beurteilung der probabilistischen Kalibrierung und Diskriminationsfähigkeit wurde zunächst der Brier Score (BS) und den daraus abgeleiteten Brier Skill Score (BSS) berechnet (vgl. Kapitel 9.1.3). Für jedes Ensemble und jeden Schwellenwert wurde die Vorhersagewahrscheinlichkeit «p» als Anteil der Ensemblemitglieder bestimmt, die die Schwelle überschreiten. Der BS beschreibt den mittleren quadratischen Fehler zwischen der prognostizierten Wahrscheinlichkeit und der tatsächlichen Beobachtung (Wilks, 2006). Als Referenz diente die klimatologische Ereigniswahrscheinlichkeit. Der daraus berechnete BSS quantifiziert den probabilistischen Skill relativ zur Klimatologie, wobei positive Werte eine Verbesserung anzeigen (Wilks, 2006).

Die Reliabilität und Resolution der Ensemblevorhersagen wurde in sogenannten Reliability-beziehungsweise Attribute-Diagrammen dargestellt. Dazu wurden die Vorhersagewahrscheinlichkeiten in Intervalle eingeteilt und der empirische Anteil tatsächlich eingetretener Ereignisse pro Intervall berechnet. Die resultierenden Punkte wurden mit Bootstrap-Konfidenzintervallen (90 %) versehen (Liechti et al., 2016). Ergänzt wurden Referenzlinien für perfekte Kalibrierung ($y = x$), No-Skill- und No-Resolution-Bereiche (Liechti et al., 2016; Wilks, 2006). Diese Diagramme ermöglichen eine direkte Einschätzung, ob die prognostizierten Wahrscheinlichkeiten tendenziell über- oder unterschätzt werden und wie stark sie zwischen verschiedenen Vorlaufzeiten variieren (Wilks, 2006).

Zur Beurteilung der Diskriminationsfähigkeit wurde für jedes meteorologische Modell, jede Station und jeden Schwellenwert eine Receiver Operating Characteristic (ROC)-Kurve berechnet. Die ROC-Kurve stellt die Beziehung zwischen der True Positive Rate (TPR) und der False Positive Rate (FPR) über alle möglichen Entscheidungsschwellen dar (Wilks, 2006). Ihre Fläche unter der Kurve (Area Under the Curve, AUC) dient als zusammenfassende Kennzahl, die unabhängig von einem einzelnen Schwellenwert interpretiert werden kann. Eine AUC von 0.5 entspricht Zufall, während Werte nahe 1 eine nahezu perfekte Diskriminationsfähigkeit anzeigen (Wilks, 2006). Die AUC-Werte wurden jahresweise und vorlaufszeitweise berechnet, wodurch die zeitliche Entwicklung der Modellleistung über die 15-jährige Untersuchungsperiode analysiert werden konnte.

3.3.4 Probabilistische Metriken für kontinuierliche Grössen

Zur quantitativen Bewertung der gesamten Ensembleverteilung wurden der Continuous Ranked Probability Score (CRPS) und der darauf basierende Continuous Ranked Probability Skill Score (CRPSS) berechnet. Der CRPS misst die mittlere Distanz zwischen der kumulativen Vorhersageverteilung und der Beobachtung in kontinuierlicher Form, er gilt als die natürliche Erweiterung des Brier Scores für stetige Grössen (Wilks, 2006). Für jedes meteorologische Modell, jede Station und jede Vorlaufszeit wurde der CRPS über alle Ensemblemitglieder bestimmt und mit einer Klimareferenz verglichen, die aus der empirischen Beobachtungsverteilung erzeugt wurde. Der daraus resultierende CRPSS zeigt den Skill relativ zur Klimatologie, wobei positive Werte eine verbesserte probabilistische Güte anzeigen. Neben Mittelwerten wurden auch Bootstrap-Konfidenzintervalle und Interquartilsabstände berechnet, um die Unsicherheiten der Schätzungen darzustellen.

Zusätzlich wurde die Ensemble-Kalibrierung mittels Rang-Histogrammen überprüft. Für jedes Ensemblemitglied wurde bestimmt, wie viele Ensemblewerte unterhalb des beobachteten Wertes lagen. Aus diesen Rängen wurde eine Häufigkeitsverteilung über alle Fälle gebildet. Eine flache Verteilung weist auf ein gut kalibriertes Ensemble hin, während U- oder umgekehrt U-förmige Formen auf Unter- beziehungsweise Überdispersion hindeuten (Wilks, 2006). Diese Analyse wurde sowohl über die gesamte Studienperiode als auch jahresweise durchgeführt, um strukturelle Veränderungen der Ensemble-Streuung im Zeitverlauf zu erkennen. Die Kombination aus CRPS/CRPSS und Rang-

Histogrammen erlaubt eine umfassende Bewertung der Ensemblequalität, einerseits hinsichtlich der Gesamtgüte der Wahrscheinlichkeitsverteilung, andererseits bezüglich der internen Konsistenz und Kalibrierung der Ensemblemitglieder.

4. Ergebnisse

4.1 Überblick und Zielsetzung

In diesem Kapitel werden die Ergebnisse der quantitativen Verifikation der neun meteorologischen Vorhersagemodelle für die Sihl vorgestellt. Die Darstellung folgt der in Kapitel 3 beschriebenen methodischen Struktur und orientiert sich an den vier Forschungsfragen der Arbeit. Im Vordergrund steht die Beurteilung der Modellgüte in Abhängigkeit von Vorlaufzeit, Schwellenwert und Modellgeneration. Besonders berücksichtigt wird die operationell relevante Abflussschwelle von 40 m³/s am Limmatwehr sowie die Pegelschwelle von 888.7 m ü. M. am Sihlsee, die den automatischen Regulierungsprozess auslöst.

Die Ergebnisse werden primär für die Station Sihlhölzli dargestellt, da sie für den Hochwasserschutz der Stadt Zürich von zentraler Bedeutung ist und aus Sicht der Endnutzer eine höhere Relevanz aufweist. Der Sihlseepegel, der im Vergleich zum Abfluss bei der Messstation Sihlhölzli deutlich weniger dynamisch auf meteorologisch-hydrologische Veränderungen reagiert, wird ergänzend analysiert, um seine Puffer- und Speicherwirkung im Gesamtsystem einzuordnen.

Zur besseren Übersichtlichkeit wurden den meteorologischen Modellen spezifische Farben zugewiesen (siehe Abbildung 7). Diese Farbcodierung bleibt über die ganze Arbeit durchgehend bestehen und dient der leichteren Unterscheidung der Modelle. Wettermodelle, die im gleichen Zeitraum aktiv waren, erhalten eine ähnliche Farbtonalität, um die Zugehörigkeit zu verdeutlichen. Eine Ausnahme bilden die Modelle Cosmo-1E, Cosmo-2E, Icon-1 und Icon-2, die in der Grafik zur Darstellung der probabilistischen Gütemasse (Kapitel 1.1) zusammengefasst werden und innerhalb der gleichen Zeitperiode gut unterscheidbar bleiben müssen.

"Cosmo-2"	=	"#e6550d"
"Cosmo-7"	=	"#fd8d3c"
"Cosmo-LEPS"	=	"#ffdd33"
"Cosmo-1"	=	"#a1d99b"
"Cosmo-E"	=	"#31a354"
"Cosmo-1E"	=	"#cc79a7"
"Cosmo-2E"	=	"#3182bd"
"Icon-1"	=	"#e6007e"
"Icon-2"	=	"#756bb1"

Abbildung 7: Farbkodierung der meteorologischen Vorhersagemodelle.

4.2 Deterministische, kontinuierliche Metriken

Die deterministische Modellgüte wurde mithilfe der in Kapitel 3.3.1 beschriebenen kontinuierlichen Metriken ME, MAE, RMSE, NSE und Bestimmtheitsmass (R^2) hinsichtlich des Abflussverlaufs und der absoluten Abweichungen zwischen Beobachtung und Vorhersage quantifiziert. Alle Werte wurden pro Modell, Vorlaufzeit und Jahr berechnet. Da die hier betrachteten Metriken deterministische Gütemasse sind, können die einzelnen Ensemblemitglieder der probabilistischen Wettermodelle nicht separat interpretiert werden. Stattdessen wurde pro Zeitschritt der Median der Ensemblemitglieder als repräsentativer Vorhersagewert herangezogen, um eine robuste und vergleichbare Bewertung zwischen deterministischen und probabilistischen Modellen zu gewährleisten (Liechti et al., 2016).

4.2.1 Allgemeine Modellgüte

Abbildung 8 zeigt die jährliche Verteilung der NSE für alle probabilistischen Wettermodelle in Abhängigkeit von der Vorlaufzeit. Jede farbige Linie repräsentiert ein Jahr im aktiven Zeitraum des jeweiligen Wettermodells. Generell ist erkennbar, dass die Vorhersagegüte mit zunehmender Vorlaufzeit abnimmt, wobei die Stärke dieses Rückgangs deutlich modellabhängig variiert. Cosmo-LEPS zeigt über alle Vorlaufzeiten hinweg eine vergleichsweise geringe Streuung der Jahres-NSE. Besonders bei 24 h und 48 h erreicht dieses Modell die höchsten Medianwerte aller Modelle und weist gleichzeitig die geringste Streuung auf. Cosmo-E und Cosmo-2E zeigen eine hohe Streuung der Jahres-NSE über den gesamten Prognosezeitraum. Der Median von Cosmo-2E nimmt zwischen 24 h und 72 h nur geringfügig ab, allerdings verschiebt sich die Verteilung nach unten, was auf eine allmähliche Abnahme der Modelleistung bei längeren Vorläufen hindeutet. Cosmo-1E und Icon-1 haben nur NSE-Werte für den 24 h Bereich, da ihre maximale Vorlaufzeit 33 h nicht überschreitet. Icon-2 zeigt bei 24 h die grösste Streuung aller Modelle, mit einem Bereich von etwa 0.6 bis -0.4. Bei 48 h verringert sich die Streuung leicht und der Median steigt geringfügig an, bleibt jedoch im negativen Bereich. Ab 72 h zeigt Icon-2 die kleinste Streuung unter allen Modellen und erzielt bei 96 h und 118 h die höchsten Median-NSE bei gleichzeitig sehr enger Verteilung. Dieses Verhalten deutet auf eine robuste und konsistente Modelleistung im längeren Vorhersagebereich hin. Zusammenfassend begünstigen kurze Vorlaufzeiten (24-48 h) insbesondere Cosmo-LEPS (hoher Median, geringe Varianz), während Icon-2 ab 72 h die stabilste und insgesamt beste Leistung zeigt. Ab 48 h fällt der Median-NSE aller Modelle unter 0.5, was die erwartete Abnahme der deterministischen Modellgüte bei längeren Prognosen bestätigt.

In Abbildung 9 ist der RMSE der probabilistischen Wettermodelle dargestellt. Insgesamt ist zu erkennen, dass die Fehler mit zunehmender Vorlaufzeit ansteigen, wobei die Unterschiede zwischen den Modellen deutlich variieren. Bei einer Vorlaufzeit von 24 h weist Icon-2 den höchsten Median-RMSE auf, dicht gefolgt von Icon-1. Cosmo-2E, Cosmo-E und Cosmo-LEPS liegen mit einem Mediawert nahe 5 deutlich darunter. Die Streuung der Fehler unterscheidet sich stark zwischen den Modellen. Während Icon-1 die engste Verteilung aufweist, zeigt Cosmo-E zwischen 24 und 72 h die grösste Spannweite. Mit zunehmender Vorlaufzeit nimmt die Medianfehlergrösse bei allen Modellen zu. Dabei stabilisiert Icon-2 seine Fehlerverteilung zunehmend und zeigt eine deutlich engere Box. Ab 72 h behält Icon-2 den niedrigsten Median-RMSE. Ab einer Vorlaufzeit von 96 h verlagert sich die grösste Streuung von Cosmo-E auf Cosmo-2E, dessen Fehlerbandbreite im längeren Prognosehorizont deutlich zunimmt. Ab 48 h zeigt Cosmo-LEPS die höchsten Median-RMSE. Zusammenfassend zeigt sich, dass Icon-2 sich als besonders robust erweist durch die engste Fehlerverteilung und niedrigsten Mediane im mehrjährigen Prognosebereich.

Nash-Sutcliffe Efficiency (NSE)

Station: Sihlhölzli | Gruppe: Probabilistisch

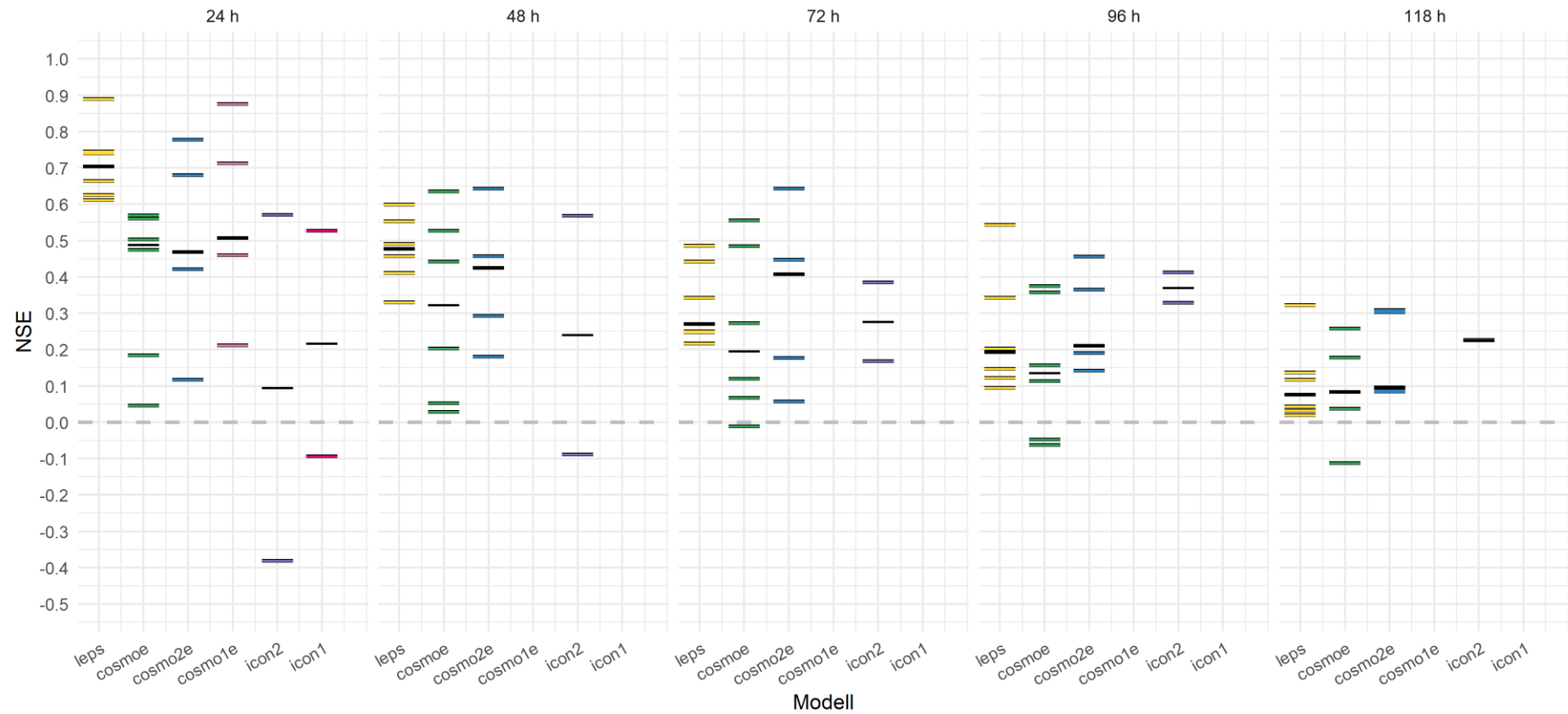


Abbildung 8: Nash–Sutcliffe Effizienz der probabilistischen Wettermodelle für die Station Sihlhölzli im Zeitraum Mai–November. Die dargestellten Werte betreffen die Vorlaufszeiten 24 h, 48 h, 72 h, 96 h und 118 h. Jede farbige Linie entspricht einem einzelnen Jahr im Zeitraum 2010–2025 und der schwarze Balken markiert den Medianwert über alle Jahre.

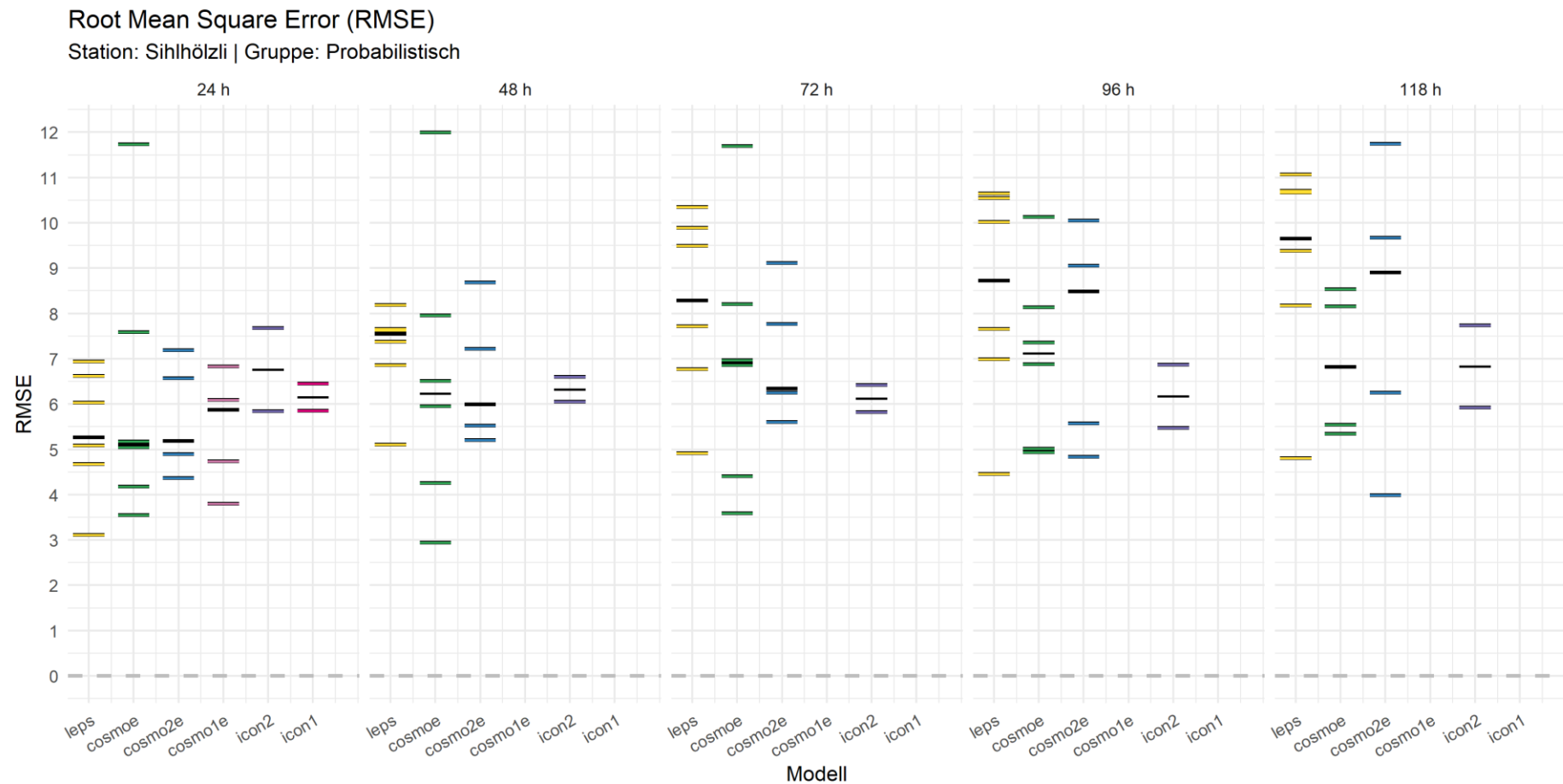


Abbildung 9: Root Mean Square Error (RMSE) der probabilistischen Wettermodelle für die Station Sihlhölzli im Zeitraum Mai–November. Die dargestellten Werte betreffen die Vorlaufzeiten 24 h, 48 h, 72 h, 96 h und 118 h. Jede farbige Linie entspricht einem einzelnen Jahr im Zeitraum 2010–2025 und der schwarze Balken markiert den Medianwert über alle Jahre.

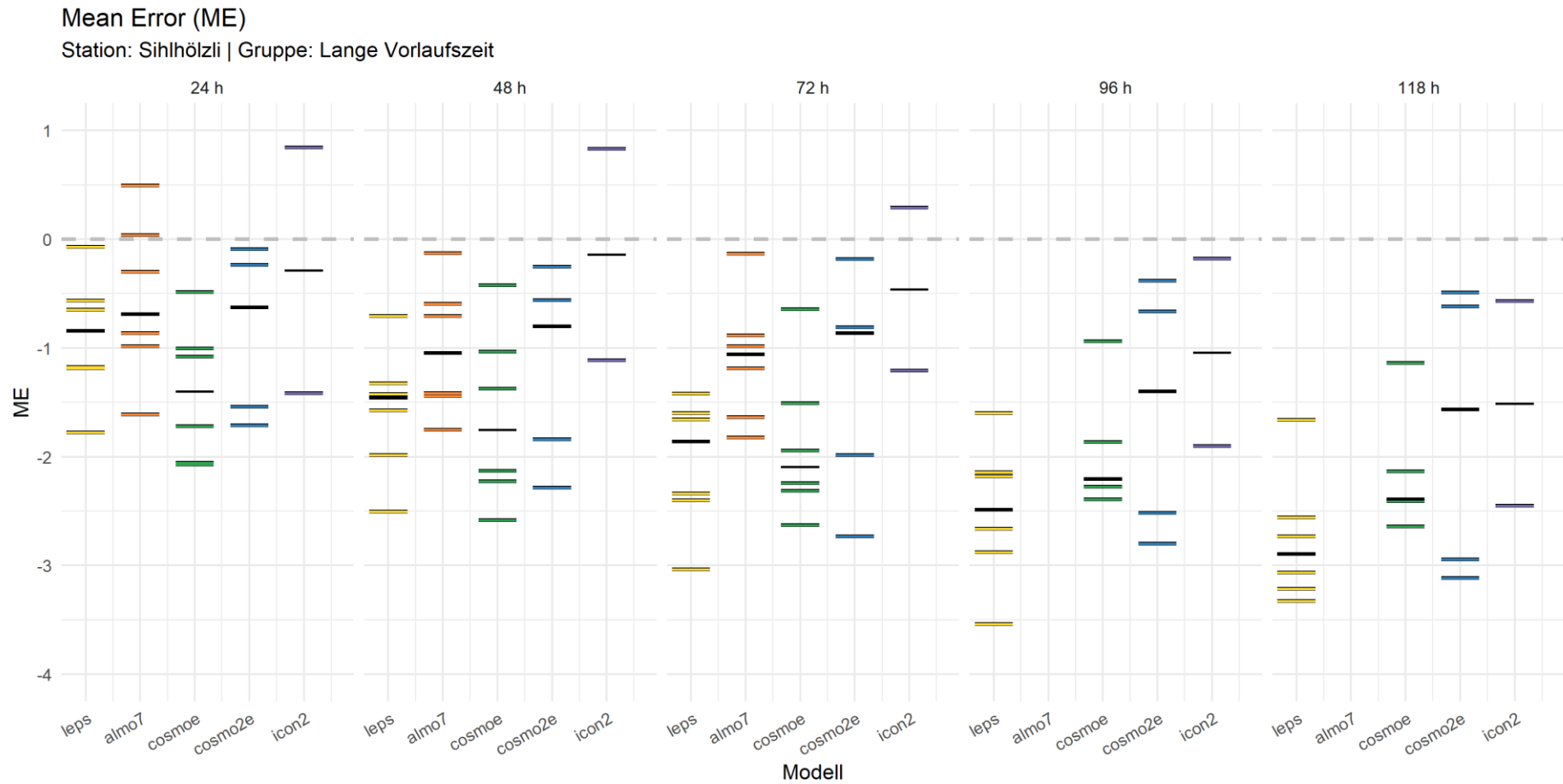


Abbildung 10: Mittlerer Fehler (ME) der Wettermodelle mit längeren Vorlaufszeit (> 33 h) für die Station Sihlhölzli im Zeitraum Mai-November. Die dargestellten Werte betreffen die Vorlaufszeiten 24 h, 48 h, 72 h, 96 h und 118 h. Jede farbige Linie entspricht einem einzelnen Jahr im Zeitraum 2010–2025 und der schwarze Balken markiert den Medianwert über alle Jahre.

4.2.2 Bias und systematische Fehler

In Abbildung 10 ist erkennbar, dass die Medianwerte des ME über alle Modelle hinweg überwiegend im negativen Bereich liegen, was auf eine tendenzielle Unterschätzung des Abflusses hinweist. Nur vereinzelt liegen Jahreswerte Werte oberhalb von Null, was auf eine leichte Überschätzung hindeutet. Bei mehreren Modellen liegen die Medianfehler sogar unter -1, was auf systematische negative Abweichungen hindeutet. Mit zunehmender Vorlaufszeit ist eine stärkere negative Verschiebung des ME zu beobachten, was bedeutet, dass die Modelle den Abfluss zunehmend stärker unterschätzen, je weiter die Prognose in die Zukunft reicht. Die Streubreite bleibt jedoch über alle Modelle hinweg weitgehend konstant, was darauf hinweist, dass die Unsicherheit der Prognosen zwar zunimmt, die Bandbreite der Fehler jedoch mit der Vorlaufzeit nicht besonders wächst. Bezüglich der Modellunterschiede zeigt sich, dass Cosmo-2E durchgehend die grösste Streuung aufweist, was auf eine variabelere Fehlerstruktur hinweist. Im Gegensatz dazu besitzt Cosmo-LEPS eine vergleichsweise kleine Streuung und eine konsistentere Fehlerverteilung, die über alle Vorlaufzeiten hinweg stabil bleibt. Icon-2 zeigt durchgehend die höchsten ME-Medianwerte (zwischen -0.1 und -1.5), was auf eine geringe Unterschätzung im Vergleich zu den anderen Modellen hinweist. Insgesamt deuten die Ergebnisse darauf hin, dass mit zunehmender Vorlaufszeit eine systematische Unterschätzung der Abflüsse auftritt, deren Stärke modellabhängig variiert.

4.2.3 Lineare Zusammenhänge

Tabelle 3 zeigt das jährlichen Bestimmtheitsmass (R^2) für alle Modelle und Vorhersagetage (Tag 1-5) an der Station Sihlhölzli. Über die gesamte Studienperiode hinweg ist ein hoher linearer Zusammenhang an kurzen Vorlaufzeiten (Tag 1) erkennbar, während diese mit zunehmender Prognosedauer deutlich abnimmt. Bei den frühen Cosmo-Modellen (Cosmo-LEPS, Cosmo-7, Cosmo-2) liegen die R^2 -Werte für Tag-1 im Bereich von 0.7 bis 0.92, was auf eine gute Reproduktion der kurzzeitigen Abflussdynamik hinweist. Bereits an Tag 2 sinken die Werte von Cosmo-LEPS auf 0.6 bis 0.7, während Cosmo-7 teilweise noch stärkere Rückgänge anzeigt. Im Vergleich dazu zeigen die Modelle Cosmo-2E und Cosmo-1E eine robustere Entwicklung über den gesamten Prognosehorizont. Die R^2 -Werte liegen an Tag 1 zwischen 0.7 und 0.95 und nehmen erst ab Tag 3-5 stärker ab. Der Rückgang zwischen Tag 1 und Tag 5 ist damit flacher als bei den älteren Modellen, was auf eine höhere Stabilität der Ensembleprognosen hindeutet. Die Icon-Modelle erreichen an Tag 1 R^2 -Werte von 0.7 bis 0.83 und sinken bis Tag 5 nur moderat auf 0.44 bis 0.65. Damit erzielen sie zwar keine höhere Anfangskorrelation als Cosmo-E oder Cosmo-2E, zeigen aber einen weniger starken Rückgang mit zunehmender Vorlaufszeit.

Tabelle 3: Jährliches Bestimmtheitsmass (R^2) zwischen modellierten und beobachteten Abflüssen an der Station Sihlhölzli im Zeitraum Mai-November, gruppiert nach Wettermodell und Vorhersagezeit (Tag 1-5). Höhere R^2 -Werte sind dunkler grün dargestellt, niedrigere heller.

		2010	2011	2012	2013	2014	2015	2016
		Mai-Nov	Mai-Nov	Mai-Nov	Mai-Nov	Mai-Nov	Mai-Nov	Mai-Nov
Cosmo Leps	Tag 1	0.7	0.82	0.92	0.92	0.78	0.8	0.8
	Tag 2	0.66	0.63	0.61	0.77	0.69	0.6	0.75
	Tag 3	0.54	0.37	0.62	0.68	0.53	0.46	0.51
	Tag 4	0.27	0.36	0.57	0.58	0.46	0.45	0.36
	Tag 5	0.27	0.25	0.38	0.58	0.29	0.35	0.32
Cosmo 7	Tag 1	0.84	0.86	0.83	0.88	0.82	0.86	0.9
	Tag 2	0.62	0.39	0.65	0.68	0.47	0.42	0.7
	Tag 3	0.46	0.24	0.46	0.55	0.29	0.31	0.5
Cosmo 2	Tag 1	0.82	0.83	0.86	0.92	0.8	0.88	0.9
		2017	2018	Mai-2019	2020			
		Mai-Nov	Nov	Mai-Nov	Mai-Nov			
Cosmo E	Tag 1	0.9	0.86	0.66	0.77	0.75		
	Tag 2	0.66	0.67	0.57	0.6	0.48		
	Tag 3	0.55	0.68	0.61	0.53	0.32		
	Tag 4	0.41	0.64	0.5	0.57	0.22		
	Tag 5	0.36	0.47	0.55	0.34	0.03		
Cosmo 1	Tag 1	0.89	0.83	0.68	0.77	0.78		
	Tag 2	0.56	0.66	0.63	0.58	0.37		
		2021	2022	2023	2024			
		Mai-Nov	Mai-Nov	Mai-Nov	Mai-Nov			
Cosmo 2E	Tag 1	0.78	0.85	0.69	0.93	0.9		
	Tag 2	0.67	0.72	0.37	0.82	0.73		
	Tag 3	0.48	0.67	0.45	0.75	0.66		
	Tag 4	0.28	0.65	0.37	0.74	0.64		
	Tag 5	0.12	0.5	0.37	0.67	0.56		
Cosmo 1E	Tag 1	0.77	0.85	0.73	0.95	0.91		
	Tag 2	0.66	0.68	0.49	0.91	0.73		
		2025						
		Mai-Aug						
Icon 2	Tag 1	0.71		0.82				
	Tag 2	0.55		0.73				
	Tag 3	0.46		0.71				
	Tag 4	0.42		0.61				
	Tag 5	0.44		0.52				
Icon 1	Tag 1	0.7		0.83				
	Tag 2	0.55		0.65				

Abbildung 11 zeigt die jährliche Entwicklung der R^2 -Werte für Tag 1, wodurch die zeitliche Stabilität der meteorologischen Modelle beurteilt werden kann. Die R^2 -Werte liegen insgesamt auf hohem Niveau (0.7-0.9), zeigen jedoch markante interannuelle Schwankungen, die modellgenerationsabhängig ausgeprägt sind. Cosmo-1 und Cosmo-E wechseln zwischen 2017 und 2019 in kurzen Abständen zwischen sehr hoher und mittlerer Modellgüte. Ein ähnliches Muster zeigen Cosmo-1E und Cosmo-2E in den Jahren 2021-2023, wo die Werte ebenfalls stark von Jahr zu Jahr streuen. Im Gegensatz dazu verlaufen die älteren deterministischen Systeme Cosmo-2 und Cosmo-7 in den Jahren 2010-2016 sehr ähnlich und weitgehend stabil. Cosmo-LEPS folgt in dieser Phase einem verwandten, aber leicht abweichenden Verlauf. Insgesamt verdeutlicht die Abbildung, dass die kurzfristige Dynamikabbildung über alle Modellgenerationen hinweg stabil bleibt, während die interannuelle Variabilität innerhalb der einzelnen Modellfamilien ein zentrales Merkmal der Vorhersagegüte darstellt.

In Abbildung 12 sind die entsprechende R^2 -Werte für Tag 5 deutlich unter denen für Tag 1 und zeigen eine ausgeprägte interannuelle Streuung. Cosmo-LEPS bewegt sich meist im Bereich von 0.25-0.40, mit einem temporären Maximum um 2013 (0.56) und einem Rückgang im Folgejahr (0.30). Cosmo-E startet 2015 mit 0.74 und fällt in den Folgejahren kontinuierlich ab, bis die Modellgüte im Jahr 2020 nahezu verschwindet. Cosmo-2E beginnt 2020 niedrig (0.28), steigt 2021 auf 0.50, fällt 2022 auf 0.37 und erreicht 2023 ein Maximum (0.66), bevor 2024 ein moderater Rückgang (0.56) einsetzt. Icon-2 liegt 2024 bei 0.44 und steigert sich 2025 auf 0.52. Sowohl Tabelle 3 als auch die Jahresverläufe der Abbildung 11 und Abbildung 12 zeigen, dass sich die kurzfristige Modellgüte über die letzten

Generationen kaum verbessert hat, längere Vorhersagen jedoch stabiler durch die geringere Streuung der R^2 -Werte über Jahre und konsistenter durch den flacheren R^2 -Abfall über die Vorlaufzeit geworden sind.

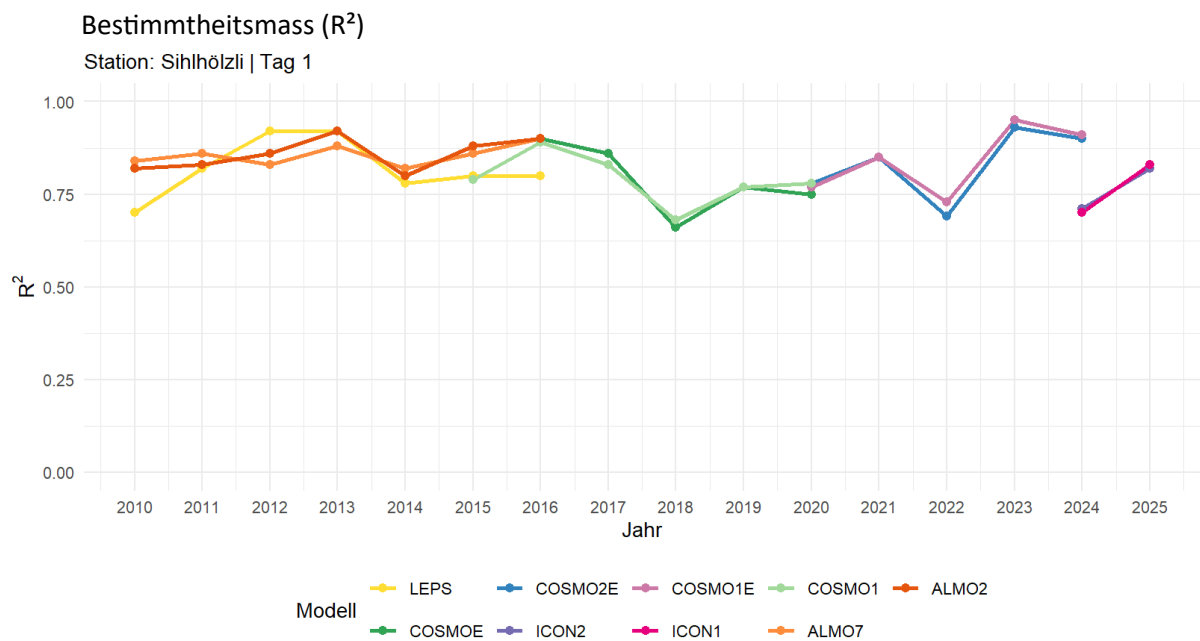


Abbildung 11: Zeitliche Entwicklung der Korrelationen des ersten Vorhersagetages pro Wettermodell und Jahr für die Station Sihlhölzli im Zeitraum Mai-November. Jedes Wettermodell ist durch eine eigene Farbe dargestellt. Während der Übergangsphasen, also zwischen der Ablösung eines alten und der Einführung eines neuen Modells, überlappen sich die meteorologischen Modelle.

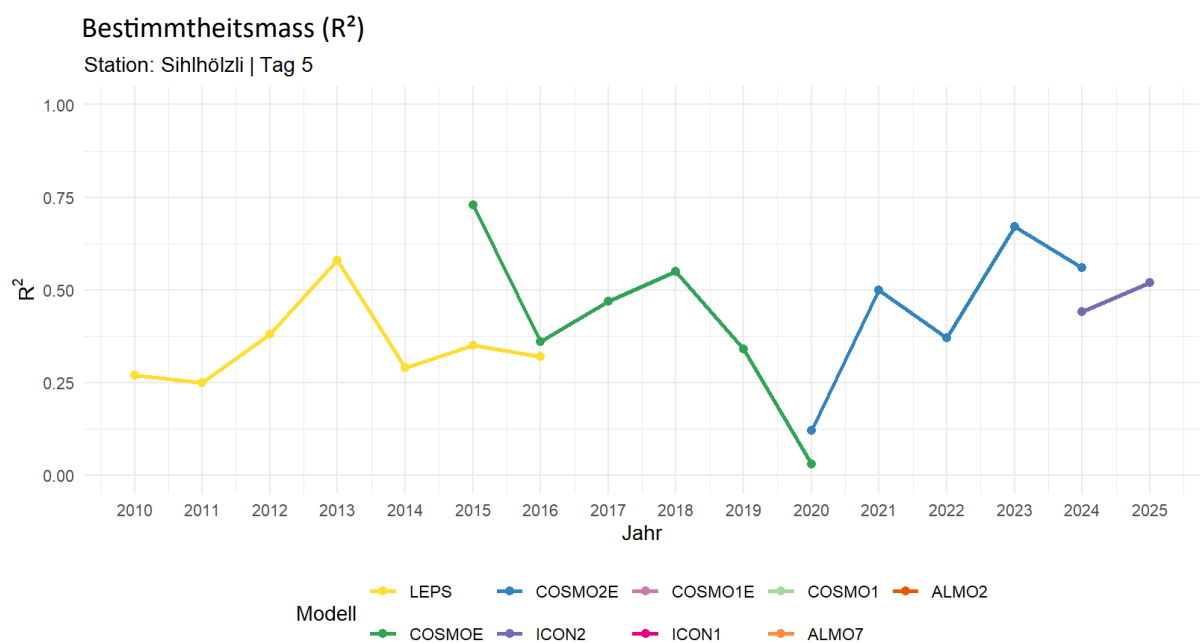


Abbildung 12: Zeitliche Entwicklung der Korrelationen des fünften Vorhersagetages pro Wettermodell und Jahr für die Station Sihlhölzli im Zeitraum Mai-November. Jedes Wettermodell ist durch eine eigene Farbe dargestellt. Während der Übergangsphasen, also zwischen der Ablösung eines alten und der Einführung eines neuen Modells, überlappen sich die meteorologischen Modelle.

4.3 Ereignisbezogene Verifikation und Schwellenanalyse

4.3.1 Allgemeine Ereignisstatistik

Die schwellenbasierte Bewertung der Vorhersagegüte wurde anhand der Metriken HSS, CSI, POD, POFD und FAR durchgeführt (vgl. Kapitel 3.3.2).

Abbildung 13 zeigt, dass mit zunehmender Schwellenhöhe und Vorlaufszeit die HSS-Werte der meteorologischen Modelle abnehmen. Für niedrige Schwellen von $10 \text{ m}^3/\text{s}$ und $20 \text{ m}^3/\text{s}$ erreichen die meisten Modelle HSS-Werte oberhalb von Null. Bereits bei $40 \text{ m}^3/\text{s}$ verlaufen jedoch mehrere Jahreslinien verschiedener Modelle auf der Nulllinie. Cosmo-1E und Icon-1 besitzen eine maximale Vorlaufszeit von rund 33 h, weshalb in der Abbildung nur die Linien für 24 h dargestellt sind. Beide Modelle zeigen, ebenso wie Icon-2, über alle Schwellen und Vorlaufszeiten hinweg eine vergleichsweise geringere Streuung der Jahres-HSS-Werte. Bei einer Schwelle von $40 \text{ m}^3/\text{s}$ weist Icon-2 im Vergleich zu Cosmo-2E durchgehend eine geringere Streuung und, mit Ausnahme der Vorlaufzeiten 24 h und 72 h, höhere Medianwerte auf. Die Medianwerte von Icon-2 liegen dabei konsistent über 0.2, variieren jedoch zwischen den Vorlaufzeiten und nehmen, anders als bei den meisten anderen Modellen, nicht kontinuierlich mit zunehmender Prognosedauer ab. Cosmo-LEPS zeigt insbesondere bei der Schwelle $10 \text{ m}^3/\text{s}$ über alle Vorlaufzeiten hinweg eine sehr geringe Streuung, die mit zunehmender Schwellenhöhe deutlich zunimmt. Ab einer Schwelle von $60 \text{ m}^3/\text{s}$ liegen nur noch die Mediane von Cosmo-1E, Icon-1 und Icon-2 im positiven Bereich, während bei $80 \text{ m}^3/\text{s}$ kein Modell mehr einen positiven Medianwert erreicht. Lediglich LEPS, Cosmo-2E und Cosmo-1E zeigen vereinzelt Jahres-HSS-Werte über Null.

In Abbildung 14 zeigt sich, dass mit zunehmender Schwellenhöhe und Vorlaufszeit eine deutliche Abnahme der CSI-Werte zu erkennen ist. Für niedrige Schwellen von $10 \text{ m}^3/\text{s}$ erreichen fast alle Modelle positive Jahreswerte, wobei Cosmo-E als einziges Modell einzelne Jahre mit Werten nahe Null aufweist. Bei einer Schwelle von $20 \text{ m}^3/\text{s}$ sinken die Medianwerte der meisten Modelle weiter ab, und die Streuung der Jahres-CSI-Werte nimmt bei mehreren Systemen bereits leicht zu. Gleichzeitig verschiebt sich die Verteilung tendenziell nach unten. Icon-1 und Icon-2 zeigen über alle Schwellen und Vorlaufzeiten hinweg eine vergleichsweise geringe Streuung. Bei $40 \text{ m}^3/\text{s}$ weist Icon-2 im Vergleich zu Cosmo-2E über alle Vorlaufzeiten hinweg eine geringere Streuung und, mit Ausnahme von 24 h, höhere Medianwerte auf. Die Medianwerte von Icon-2 verbleiben insgesamt im positiven Bereich, variieren jedoch zwischen den Vorlaufzeiten und nehmen, anders als bei den übrigen Modellen, nicht gleichmäßig mit zunehmender Prognosedauer ab. Ab einer Schwelle von $60 \text{ m}^3/\text{s}$ liegen nur noch die Medianwerte von Cosmo-1E, Icon-1 und Icon-2 im positiven Bereich, während bei $80 \text{ m}^3/\text{s}$ kein Modell mehr einen positiven Medianwert erreicht. Vereinzelt treten bei LEPS, Cosmo-2E und Cosmo-1E noch positive Jahreswerte auf.

Heidke Skill Score (HSS)
Station: Sihlhölzli | Gruppe: Probabilistisch

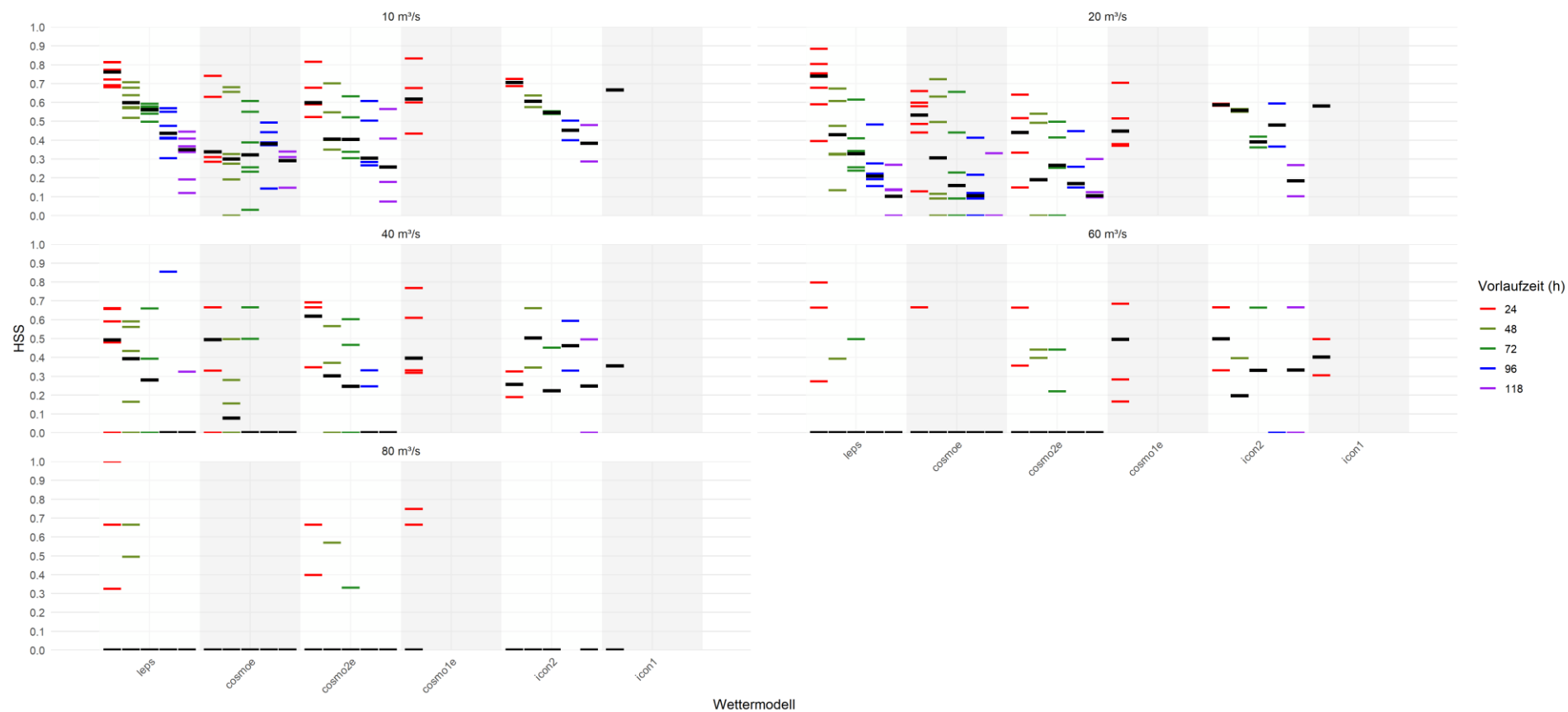


Abbildung 13: Heidke Skill Score (HSS) der probabilistischen Wettermodelle, facettiert nach Schwellwert und Darstellung der Vorlaufzeiten als Linien für die Station Sihlhölzli im Zeitraum Mai-November. Jede farbige Linie entspricht einem einzelnen Jahr im Zeitraum 2010-2025 und der schwarze Balken markiert den Medianwert über alle Jahre.

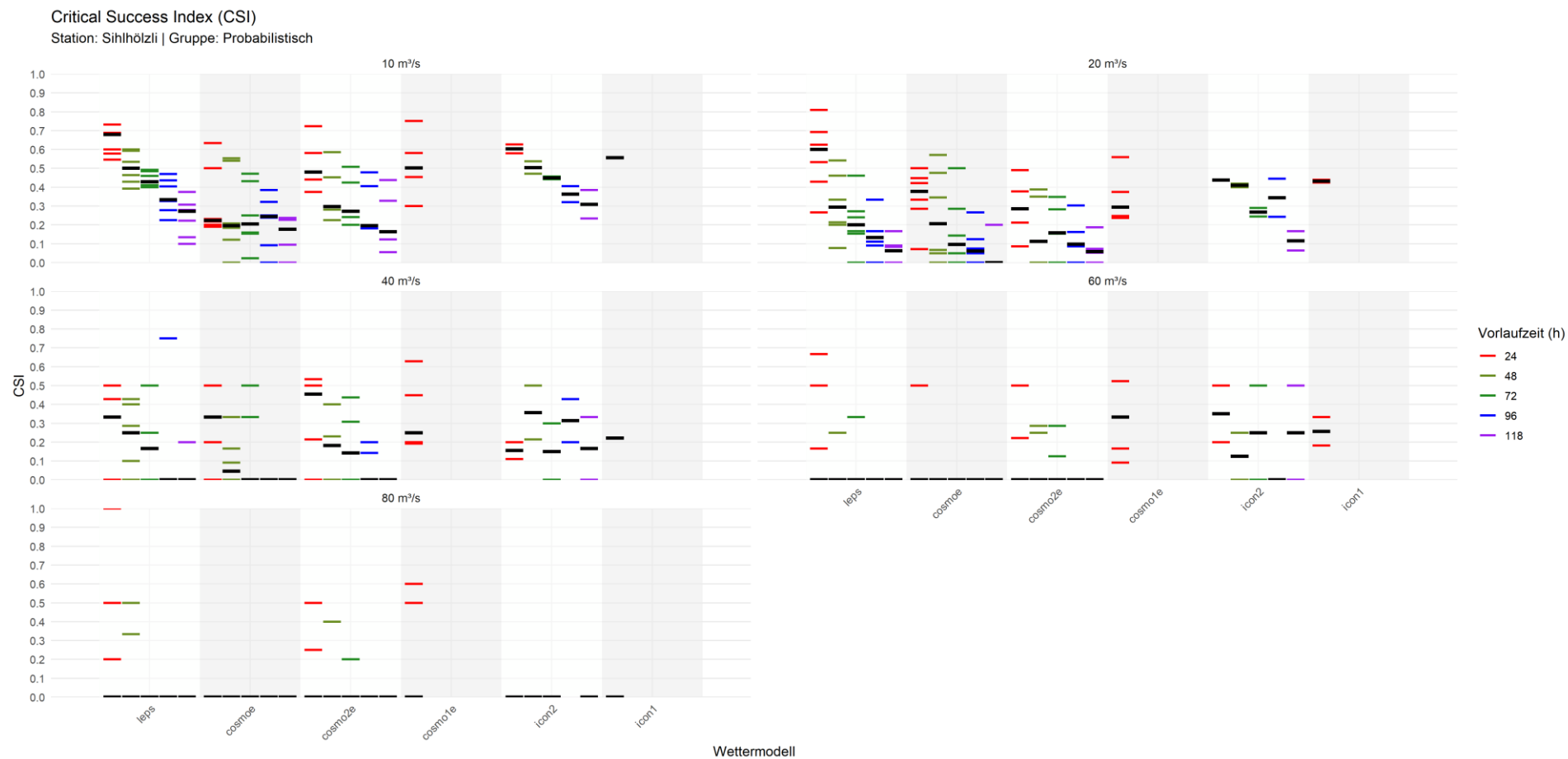


Abbildung 14: Critical Success Index (CSI) der probabilistischen Wettermodelle, facettiert nach Schwellwert und Darstellung der Vorlaufzeiten als Linien für die Station Sihlhölzli im Zeitraum Mai-November. Jede farbige Linie entspricht einem einzelnen Jahr im Zeitraum 2010-2025 und der schwarze Balken markiert den Medianwert über alle Jahre.

4.4 Probabilistische Metriken für binäre Ereignisse

Die probabilistische Bewertung der binären Ereignisse basiert auf dem Brier Score (BS), dem Brier Skill Score (BSS) sowie den aus ROC-Kurven abgeleiteten AUC-Werten (vgl. Kapitel 3.3.3).

4.4.1 Gesamtgüte der probabilistischen Vorhersagen

In Abbildung 15 ist der BSS der Modelle Icon-1 und Icon-2 über die gesamte Vorlaufszeit hinweg dargestellt. Für die niedrigen Schwellen (10 und 20 m³/s) zeigen beide Modelle durchgehend positive BSS-Werte, wobei der Median von Icon-2 bei 20 m³/s und einer Vorlaufszeit von etwa 36 h in den negativen Bereich fällt. Die beiden Modelle unterscheiden sich insgesamt nur geringfügig voneinander, ebenso die jeweiligen Medianverläufe, die über weite Teile des Prognosehorizonts auf einem ähnlichen Niveau verbleiben. Insgesamt ist ein leichter, aber konsistenter Abwärtstrend der BSS-Werte mit zunehmender Vorlaufszeit zu erkennen. Ab der Schwelle 40 m³/s beginnen die Werte stärker zu streuen und nehmen mit zunehmender Vorlaufszeit deutlich ab. Der Median von Icon-2 fällt bereits vor 12 h Vorlaufszeit in den negativen Bereich, gefolgt vom Median von Icon-1 ab etwa 16 h Vorlaufszeit. Die Verläufe der beiden Modelle sinken schrittweise unter die Nulllinie, wobei die BSS-Werte > -0.5 verbleiben. In diesem Bereich deuten negative Werte punktuell auf Über- oder Unterschätzungen der Eintrittswahrscheinlichkeit seltener Ereignisse hin. Bei den hohen Schwellen (≥ 60 m³/s) zeigt sich ein deutlich abgehacktes Bild mit einzelnen, isolierten Punkten und fehlenden Liniensegmenten. Diese Lücken entstehen dadurch, dass für die höchsten Schwellen keine oder nur sehr wenige beobachtete Überschreitungen vorliegen, wodurch der BSS für einzelne Vorlaufzeiten nicht berechnet werden kann.

Zum Vergleich zeigt Abbildung 16 den BSS der Modelle Cosmo-1E und Cosmo-2E sowie deren jeweiligen Medianverläufe. Für niedrige und mittlere Schwellen (10-40 m³/s) weisen beide Modelle durchgehend positive BSS-Werte auf, was auf eine insgesamt gute probabilistische Kalibrierung hinweist. Mit zunehmender Vorlaufszeit nimmt der BSS allmählich ab, bleibt jedoch überwiegend positiv. Beide Modelle zeigen dabei leicht höhere Werte als ihre jeweiligen Medianverläufe. Ab einer Schwelle von 60 m³/s erreicht der Median von Cosmo-2E vereinzelt negative Werte und verläuft um die Nulllinie. Bei 80 m³/s liegen die anfänglichen BSS-Werte deutlich tiefer als bei den niedrigeren Schwellen, teilweise bereits im negativen Bereich. Sie steigen bis zu einer Vorlaufszeit von rund 15 h an, bevor Cosmo-2E und sein Median erneut absinken, während Cosmo-1E über denselben Zeitraum weitgehend stabil bleibt. Bei einer Vorlaufszeit von etwa 54 h erreicht Cosmo-2E einen BSS-Wert von Null und schwankt danach im Bereich zwischen 0 und 0.25.

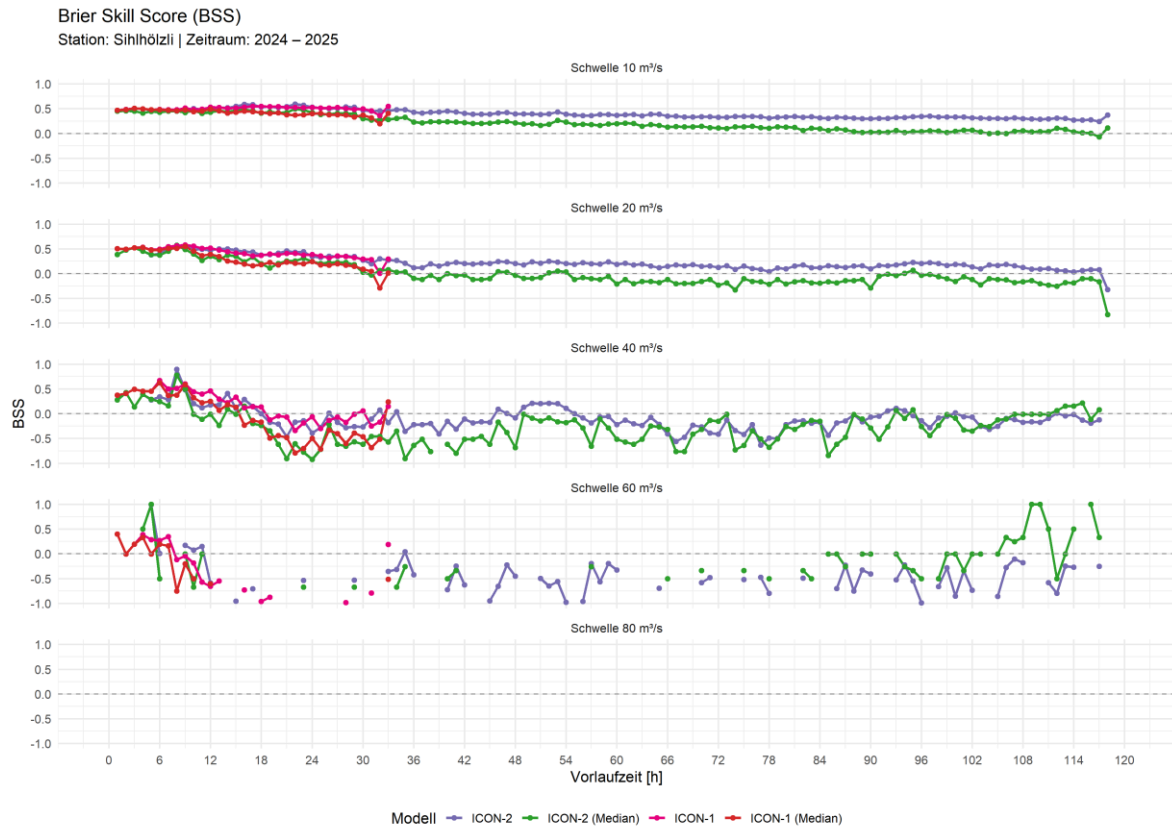


Abbildung 15: Brier Skill Score (BSS) der meteorologischen Modelle Icon-1 und Icon-2 für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2024-2025.

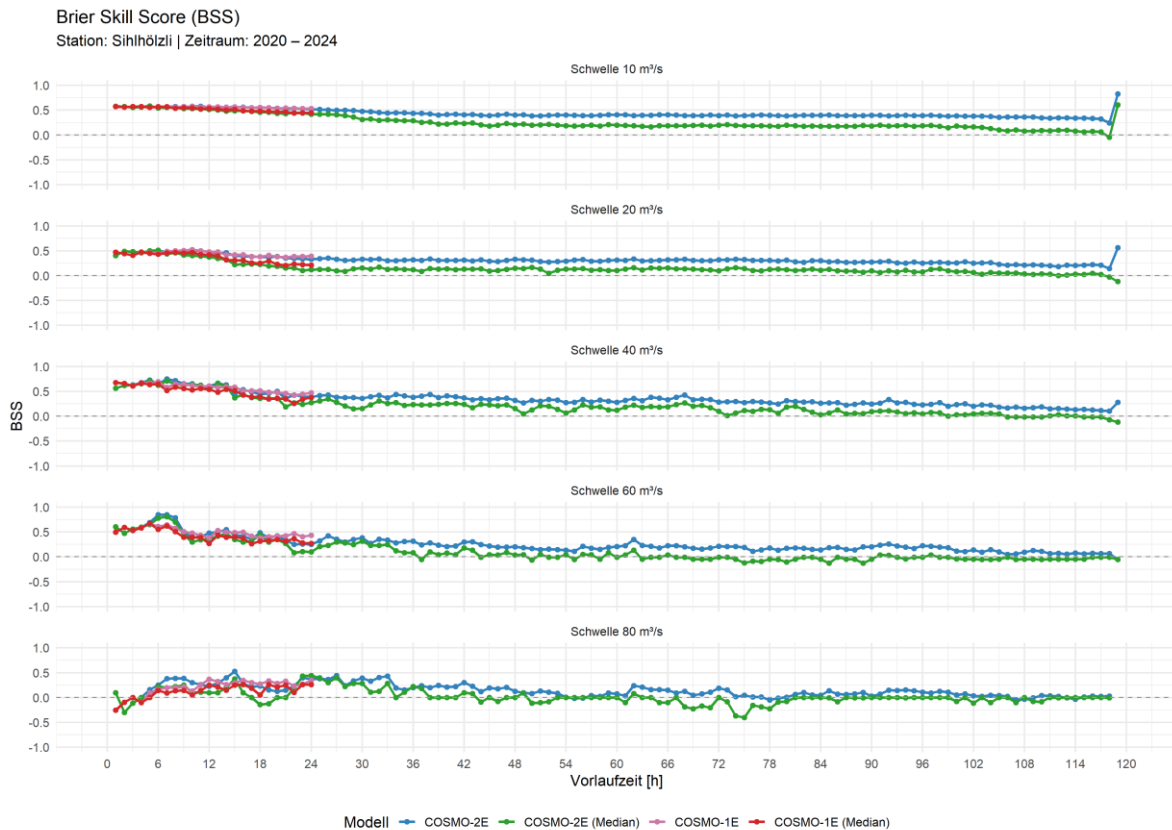


Abbildung 16: Brier Skill Score (BSS) der meteorologischen Modelle Cosmo-1E und Cosmo-2E für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2020-2024.

4.4.2 Diskriminationsfähigkeit der meteorologischen Modelle

Abbildung 17 zeigt die ROC-Kurven des Modells Icon-2 für die Station Sihlhölzli, getrennt nach den vier Abflussschwellen (10, 20, 40 und 60 m³/s). Für die höchste Schwelle (80 m³/s) wurde keine Facette dargestellt, da aufgrund der geringen Anzahl an beobachteten und nicht beobachteten Ereignissen oberhalb dieser Schwelle keine gültige ROC-Kurve berechnet werden konnte. Die Berechnung wurde deshalb im Auswertungsskript automatisch übersprungen. Für die niedrigen Schwellen (10 und 20 m³/s) verlaufen die Kurven glatt und stark konvex zur linken oberen Ecke, was auf eine hohe Diskriminationsfähigkeit des Modells hinweist. Die Linien für kürzere Vorlaufzeiten (24 h und 48 h) liegen dabei konsistent am nächsten an der idealen Ecke (0, 1) und erreichen somit die höchste Trefferwahrscheinlichkeit bei gleichzeitig geringer Fehlalarmrate. Mit zunehmender Vorlaufzeit (72 h, 96 h, 118 h) flachen die Kurven ab, was eine Abnahme der Trennschärfe zwischen Ereignis und Nicht-Ereignis widerspiegelt. Bei den höheren Schwellen (40 und 60 m³/s) werden die Kurven zunehmend stufenförmig und unregelmässig, was auf eine geringe Anzahl verfügbarer Ereignisse zurückzuführen ist. Dadurch werden die Schwellenbereiche diskret abgebildet, und kleine Änderungen in der Wahrscheinlichkeit führen zu abrupten Sprüngen in der True-/False-Positive-Rate. Besonders bei 60 m³/s zeigt sich dieses Verhalten deutlich, was auf eine statistische Instabilität aufgrund seltener Ereignisse hindeutet. Ein interessantes Muster zeigt sich bei der Schwelle 40 m³/s: Hier verlaufen die Kurven der ersten drei Vorlaufzeiten (24 h, 48 h, 72 h) teilweise deckungsgleich und in einzelnen Bereichen übertrifft die 48-h- oder 72-h-Kurve sogar die 24-h-Kurve leicht. Dies deutet darauf hin, dass die Diskriminationsfähigkeit des Modells bei moderaten Hochwasserereignissen über den ersten drei Prognosetage hinweg relativ stabil bleibt. Zum Vergleich zeigt Abbildung 18 die ROC-Kurven des Modells Cosmo-2E. Für die Schwellen von 10 bis 40 m³/s verlaufen die Kurven gleichmässig und glatt nahe der idealen oberen linken Ecke und weisen damit auf eine sehr hohe Diskriminationsfähigkeit hin. Bei einer Vorlaufzeit von 119 h flacht die Kurve leicht ab, bleibt jedoch deutlich über der Diagonalen und weiterhin stark konvex. Bei den hohen Schwellen von 60 m³/s und 80 m³/s treten zwar ebenfalls stufenartige Verläufe auf, doch liegen die Linien insgesamt näher an der oberen linken Ecke als bei Icon-2. Besonders bei 60 m³/s zeigt Cosmo-2E deutlich glattere und stärker konvexe Kurven, auch bei langen Vorlaufzeiten, was auf eine robustere und stabilere Diskriminationsleistung im Extrembereich hinweist.

Um die zeitliche Stabilität der modell- und schwellenabhängigen Diskriminationsfähigkeit über verschiedene Vorlaufzeiten hinweg zu beurteilen, zeigen die Abbildung 19 und Abbildung 20 die jährliche Entwicklung der AUC für alle Modelle bei einer Vorlaufzeit von 24 und 118 h und der Schwelle von 40 m³/s. Über den gesamten Zeitraum hinweg liegen die AUC-Werte aller Modelle bei Vorlaufzeit 24 h deutlich über 0.88 (mit Ausnahme von Cosmo-E 2019). Besonders Cosmo-LEPS und Cosmo-7 zeigen über die Jahre hinweg eine stabile Güte mit AUC-Werten nahe 1.0. Die nachfolgenden Cosmo- und Icon-Modelle erreichen ebenfalls sehr hohe Werte, weisen jedoch eine leicht stärkere interannuelle Variabilität auf. Einzelne Schwankungen, etwa ein Rückgang der Werte bei Cosmo-E um 2019 oder bei Cosmo-2E um 2021 und 2022, deuten auf Jahre mit besonders anspruchsvollen hydrometeorologischen Bedingungen hin, bei denen die Trennschärfe der Modelle kurzfristig abnahm. Die Werte liegen durchweg weit über dem Zufallsniveau (AUC = 0.5), was eine robuste Fähigkeit aller Modelle zur Erkennung von Ereignissen oberhalb der Schwelle von 40 m³/s bestätigt. Auch bei einer Vorlaufzeit von 118 h sind Unterschiede zwischen den Modellgenerationen erkennbar. Cosmo-Leps steigt zwischen 2010 und 2013 von 0.75 auf 1 an, um danach steil auf 0.65 abzufallen. In den anschliessenden Jahren bleibt der Wert auf diesem tieferen Niveau aber weitgehend konstant. Auch Cosmo-E steigt zwischen 2016 und 2018 von einer ähnlichen Anfangsgüte (AUC ≈ 0.75) steil auf 1 an, aber fällt bis 2020 steil auf unter 0.5, also unter dem Zufallsniveau. Cosmo-2E zeigt zwischen 2020 und 2024 eine Steigung von etwa 0.8 auf rund 0.9. Das jüngste Modell Icon-2 startet 2024 bei einem hohen

AUC-Wert von rund 0.88 und fällt 2025 leicht auf 0.8, verbleibt damit jedoch deutlich über dem Zufallsniveau.

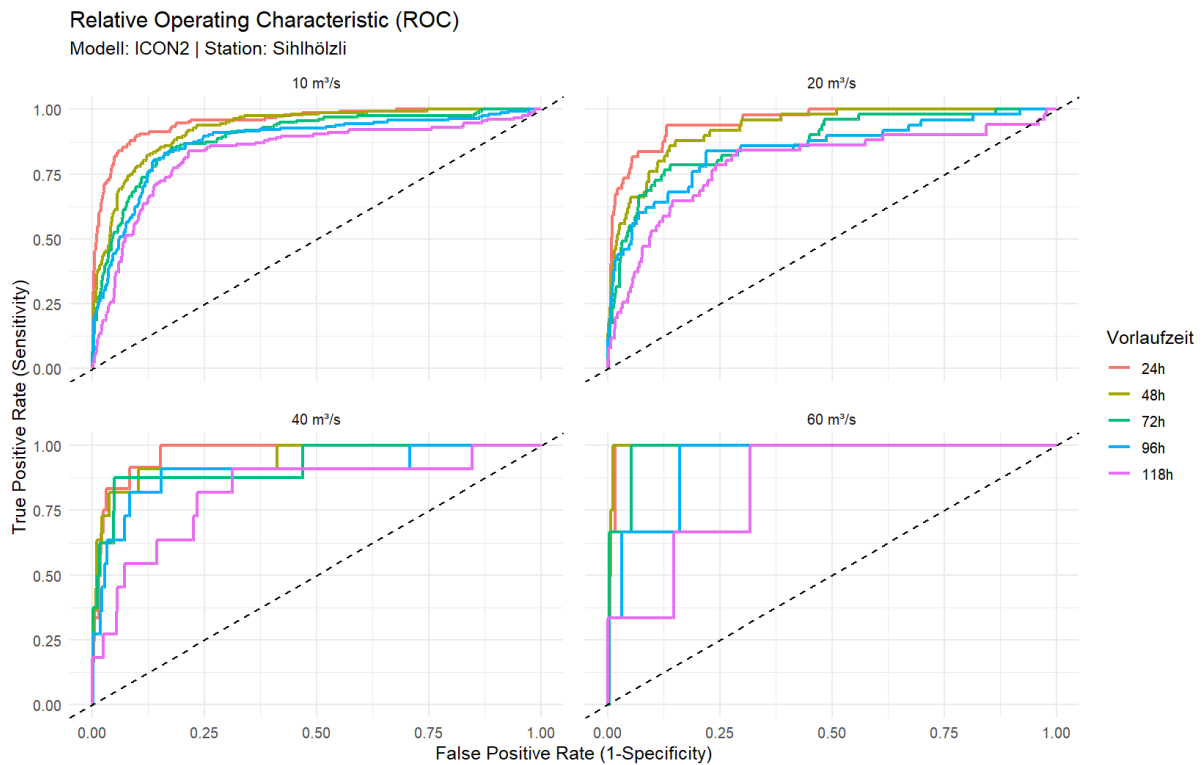


Abbildung 17: Receiver Operating Characteristic (ROC) des meteorologischen Modells Icon-2 für verschiedene Abflussschwellen und Vorlaufzeiten an der Station Sihlhölzli im Zeitraum Mai-November 2024-2025. Die farbigen Kurven zeigen den Zusammenhang zwischen Trefferquote (True Positive Rate) und Fehlalarmrate (False Positive Rate).

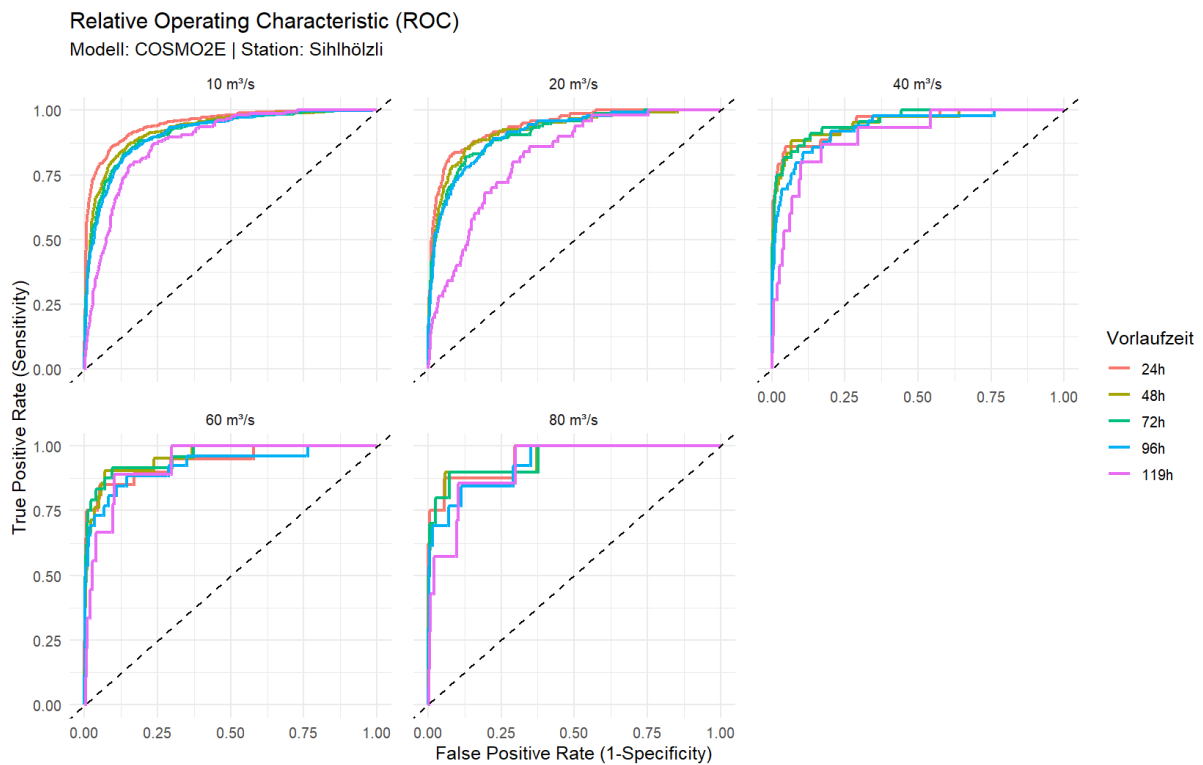


Abbildung 18: Receiver Operating Characteristic (ROC) des meteorologischen Modells Cosmo-2E für verschiedene Abflussschwellen und Vorlaufzeiten an der Station Sihlhölzli im Zeitraum Mai-November 2020-2024. Die farbigen Kurven zeigen den Zusammenhang zwischen Trefferquote (True Positive Rate) und Fehlalarmrate (False Positive Rate).

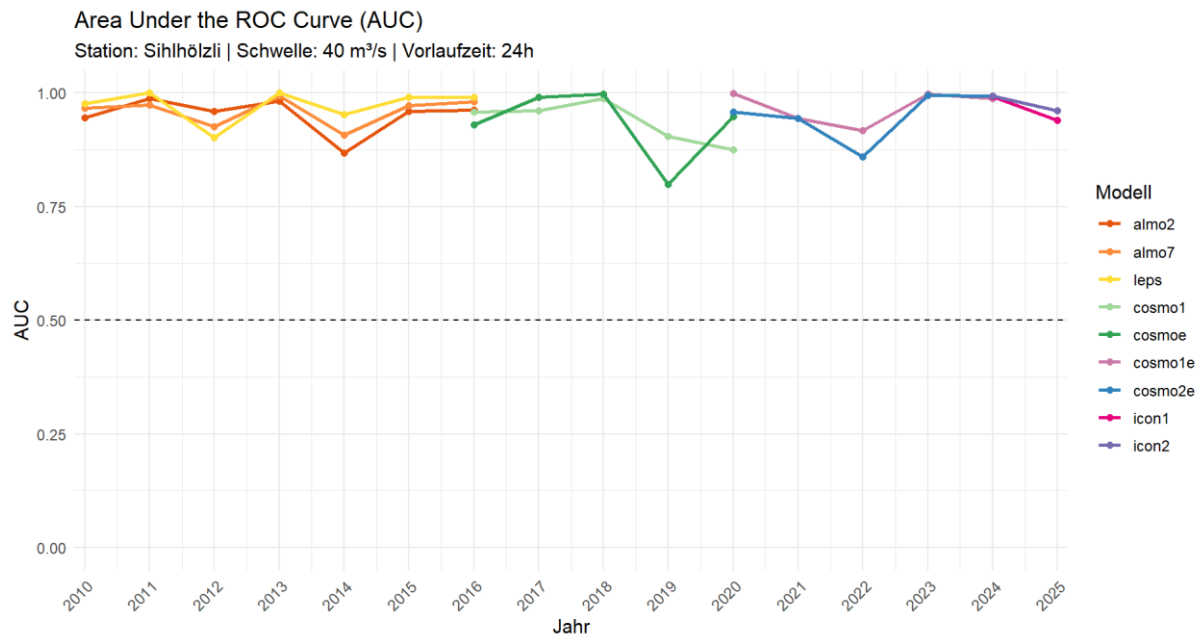


Abbildung 19: Area under the Curve (AUC) der meteorologischen Modelle für eine Vorlaufzeit von 24 Stunden und eine Abflussschwelle von 40 m³/s an der Station Sihlhölzli im Zeitraum Mai bis November 2010–2025. Jedes Wettermodell ist durch eine eigene Farbe dargestellt. Während der Übergangsphasen, also zwischen der Ablösung eines alten und der Einführung eines neuen Modells, überlappen sich die jeweiligen AUC-Werte.

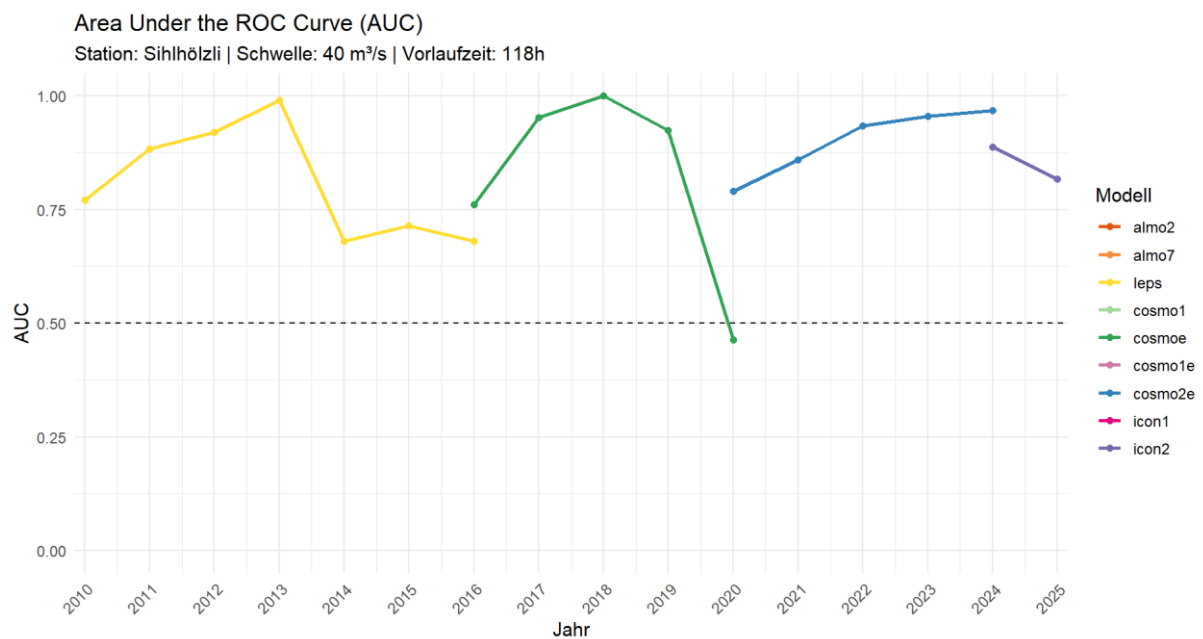


Abbildung 20: Area under the Curve (AUC) der meteorologischen Modelle für eine Vorlaufzeit von 118 Stunden und eine Abflussschwelle von 40 m³/s an der Station Sihlhölzli im Zeitraum Mai bis November 2010–2025. Jedes Wettermodell ist durch eine eigene Farbe dargestellt. Während der Übergangsphasen, also zwischen der Ablösung eines alten und der Einführung eines neuen Modells, überlappen sich die jeweiligen AUC-Werte.

4.4.3 Zuverlässigkeit der Wahrscheinlichkeiten

In Abbildung 21 und Abbildung 22 sind die Attributdiagramme der meteorologischen Modelle Icon-2 und Cosmo-2E für die Abflussschwelle 40 m³/s ersichtlich. Für Icon-2 liegen die Punkte der Vorlaufszeitgruppen 1-24 h, 25-48 h, 49-72 h und 73-96 h durchgehend unterhalb der Perfect-Reliability-Linie ($y=x$). In der Gruppe 97-120 h befinden sich Punkte oberhalb der Diagonalen, jedoch erst bei Vorhersagewahrscheinlichkeiten $p > 0.7$. In den Gruppen 1-24 h und 25-48 h liegen alle Punkte mit $p < 0.75$ unterhalb der No-Skill-Linie, während Punkte mit $p \geq 0.75$ oberhalb der No-Skill-Linie, jedoch weiterhin unterhalb der Diagonalen liegen. Die 90 %-Konfidenzintervalle unterscheiden sich in ihrer Länge zwischen den Wahrscheinlichkeitsklassen. Sie sind überwiegend ab $p \approx 0.5$ sichtbar, darunter sind sie meist zu kurz. Die kleinen Zahlen neben den Punkten kennzeichnen den relativen Anteil der tatsächlich eingetretenen Ereignisse pro Intervall und variieren zwischen den Vorlaufzeiten und Wahrscheinlichkeitsklassen. Sie treten teils bei hohen, teils bei niedrigen Vorhersagewahrscheinlichkeiten als Maxima auf, ohne erkennbares Muster.

Für Cosmo-2E folgen die Punkte über weite Bereiche der 1:1-Diagonalen, wobei die Abweichungen mit zunehmender Vorlaufzeit zunehmen. In den Gruppen 1-24 h und 25-48 h liegen alle Punkte nahe der Diagonalen, bei 1-24 h treten ab einer Vorhersagewahrscheinlichkeit von $p = 0.6$ geringe Abweichungen unterhalb der Diagonalen auf. Ab der Gruppe 25-48 h nehmen die Abweichungen oberhalb der Diagonalen zu, beginnend etwa ab $p = 0.25-0.3$. In den Gruppen 73-96 h und 97-120 h liegen die Punkte für $p = 0.25-0.6$ überwiegend oberhalb, bei $p > 0.6$ hingegen unterhalb der Diagonalen. Die Konfidenzintervalle variieren in ihrer Länge und sind insbesondere in den Gruppen 73-96 h und 97-120 h ab $p \approx 0.6$ deutlich breiter. Bei 97-120 h erstrecken sich einzelne Intervalle sogar von oberhalb der Diagonalen bis unterhalb der No-Skill-Linie. Über alle Vorlaufzeiten hinweg befindet sich ein Grossteil der Daten in der niedrigsten Wahrscheinlichkeitsklasse, während die übrigen Klassen nur wenige Datenpunkte enthalten.

Attributes Diagram

Modell: ICON-2 | Station: Sihlhölzli | Schwellwert: 40 m³/s

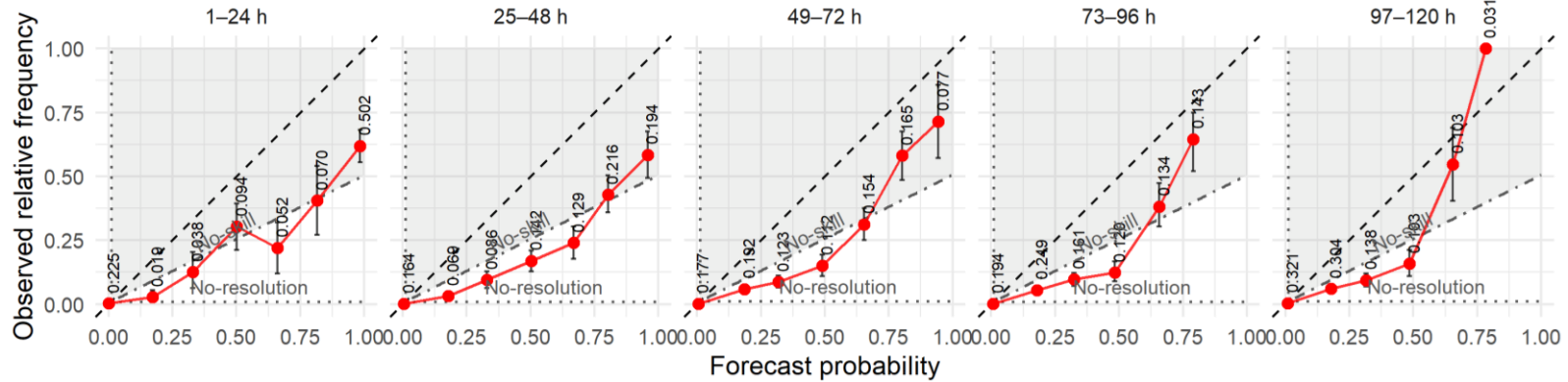


Abbildung 21: Attribut Diagramm des meteorologischen Modells Icon-2 für die Abflussschwelle von 40 m³/s und verschiedene Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2024-2025.

Attributes Diagram

Modell: Cosmo-2E | Station: Sihlhölzli | Schwellwert: 40 m³/s

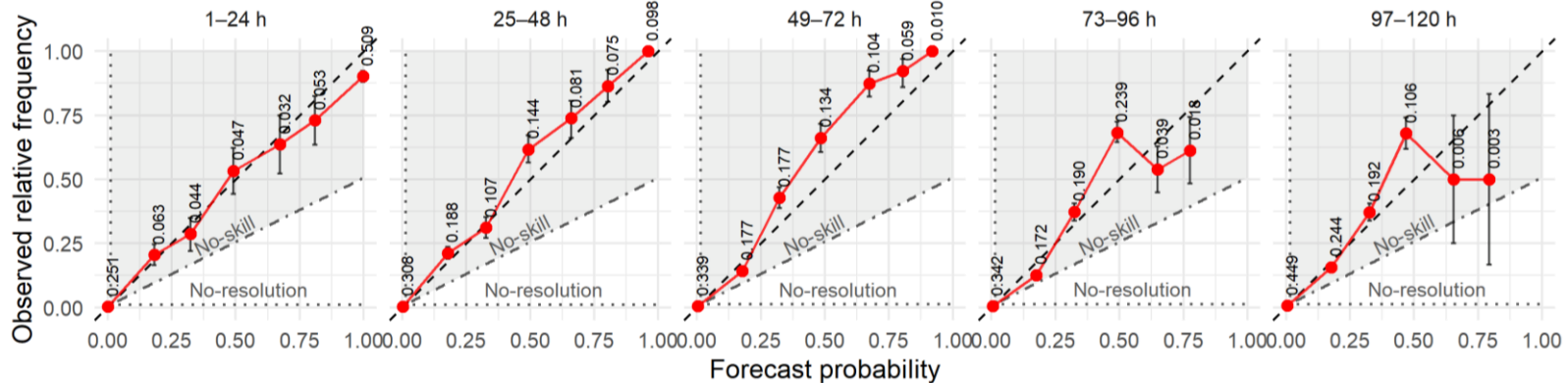


Abbildung 22: Attribut Diagramm des meteorologischen Modells Cosmo-2E für die Abflussschwelle von 40 m³/s und verschiedene Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2020-2024.

4.5 Probabilistische Metriken für kontinuierliche Grössen

Für die probabilistische Bewertung kontinuierlicher Grössen wurden der Continuous Ranked Probability Score (CRPS) und der Continuous Ranked Probability Skill Score (CRPSS) berechnet (vgl. Kapitel 3.3.4).

4.5.1 Gesamtgüte der Ensembleverteilung

In Abbildung 23 und Abbildung 24 ist der mittlere CRPS-Wert der meteorologischen Modelle Icon-1, Icon-2, Cosmo-1E und Cosmo-2E über alle Fälle dargestellt. Die schattierten Bänder zeigen die zugehörigen 90 %-Konfidenzintervalle und verdeutlichen die Unsicherheit des CRPS-Mittelwerts. Schmale Bänder deuten auf eine präzisere Schätzung hin (viele Fälle und/oder geringe Streuung), breite Bänder entsprechend auf eine unsicherere Schätzung (wenige Fälle und/oder grosse Streuung). Für Icon-1 und Icon-2 beginnen die CRPS-Werte bei einer Vorlaufszeit von 0 h bei etwa 1.6 und nehmen mit zunehmender Vorlaufszeit zu. Bei 24 h liegen beide Wettermodelle bei rund 2.0, bei 48 h erreicht Icon-2 einen CRPS-Wert von etwa 2.2, bei 72 h und 96 h rund 2.3, und bei 118 h steigt der Wert auf etwa 2.5. Die Verläufe beider Wettermodelle zeigen über die gemeinsamen Vorlaufszeiten hinweg nahezu deckungsgleiche Werte. Lediglich bei Icon-2 sind geringfügige wellenförmige Schwankungen erkennbar, während Icon-1 einen gleichmässig ansteigenden Verlauf ohne sichtbare Variation aufweist. Die wellenförmige Schwankungen bleiben bei Icon-2 über die ganze Vorlaufzeitperiode bestehen. Die Konfidenzintervalle beider Wettermodelle sind in den ersten Stunden (bis ca. 10 h) am schmalsten, nehmen bis etwa 24 h leicht zu und bleiben ab dann weitgehend konstant.

Die CRPS-Werte von Cosmo-1E und Cosmo-2E beginnen bei einer Vorlaufszeit von 0 h bei etwa 1.7. In den ersten 10 Stunden sinken die Werte leicht ab, bevor sie mit zunehmender Vorlaufszeit wieder ansteigen. Bei Cosmo-2E steigt der CRPS bis 118 h auf rund 2.4 und erreicht bei 119 h etwa 2.6, wobei der letzte Zeitschritt einen deutlicheren Sprung aufweist. Bis 24 h liegen die Werte von Cosmo-1E leicht über jenen von Cosmo-2E. Im ganzen Verlauf zeigt Cosmo-2E wellenförmige, geringfügige Schwankungen, während der Verlauf von Cosmo-1E gleichmässiger ansteigt. Die Konfidenzintervalle beider Wettermodelle sind im Vergleich zu den Icon-Modellen insgesamt schmäler und zeigen über alle Vorlaufszeiten hinweg eine weitgehend konstante Breite.

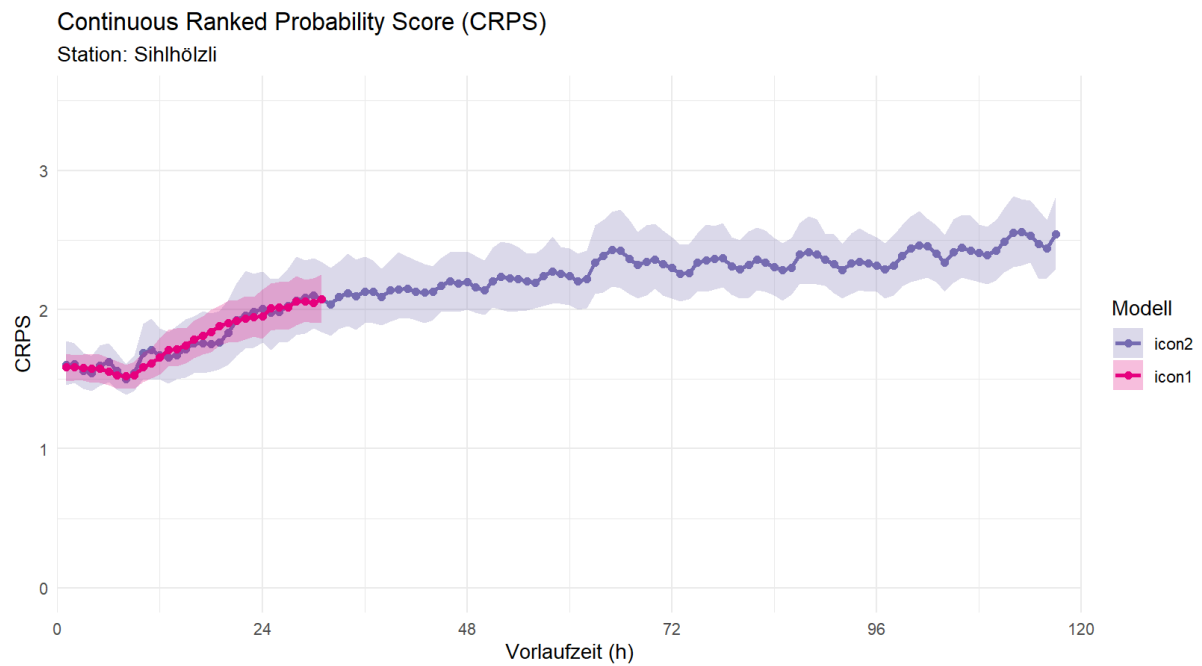


Abbildung 23: Continuous Ranked Probability Score (CRPS) der meteorologischen Modelle Icon-1 und Icon-2 in Abhängigkeit der Vorlaufzeit an der Station Sihlhölzli im Zeitraum Mai bis November 2024-2025.

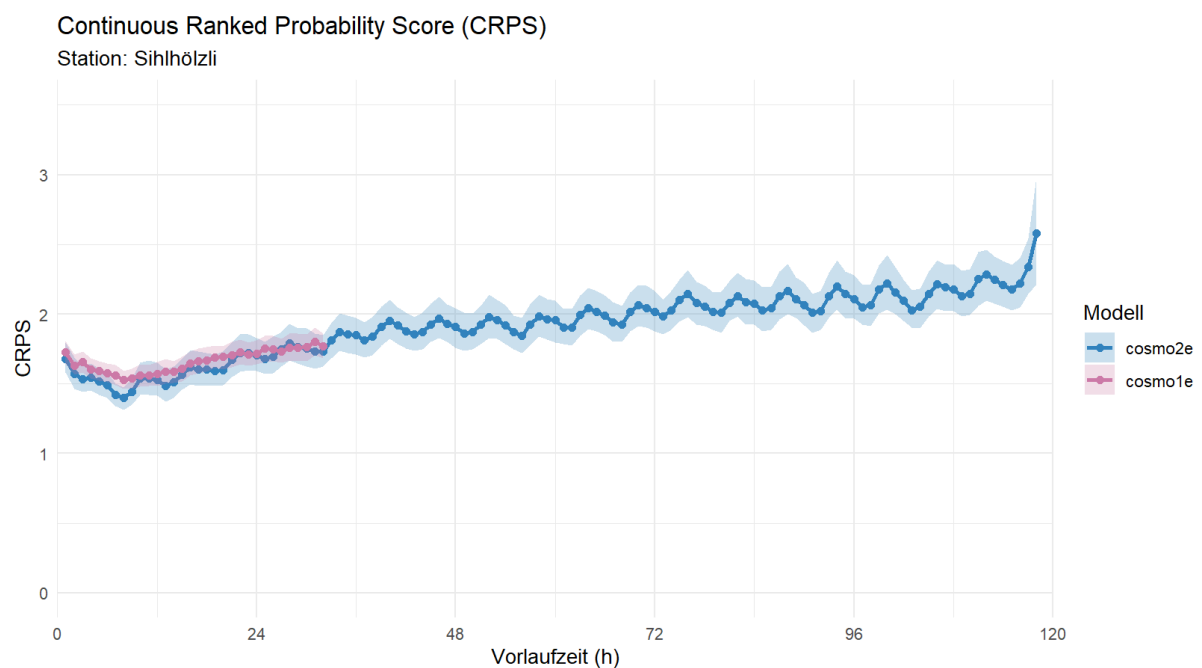


Abbildung 24: Continuous Ranked Probability Score (CRPS) der meteorologischen Modelle Cosmo-1E und Cosmo-2 E in Abhängigkeit der Vorlaufzeit an der Station Sihlhölzli im Zeitraum Mai bis November 2020-2024

4.5.2 Zeitliche Entwicklung der probabilistischen Modellgüte

In Abbildung 25 und Abbildung 26 ist der zeitliche Verlauf des mittleren CRPS aller meteorologischen Modelle für eine Vorlaufszeit von 24 h (Abbildung 25) beziehungsweise 120 h (Abbildung 26) dargestellt. Die schattierten Bänder zeigen die zugehörigen 90 %-Konfidenzintervalle und verdeutlichen die Unsicherheit des jeweiligen Jahresmittelwerts. Bei einer Vorlaufszeit von 24 h zeigt Cosmo-LEPS zu Beginn den höchsten CRPS-Wert mit etwa 2.7, gefolgt von einem zickzackförmigen Abfall bis zum Jahr 2016, wo ein Wert von rund 2.1 erreicht wird. Das Konfidenzintervall ist zu Beginn am breitesten und wird im Verlauf enger. Cosmo-E startet mit dem breitesten Konfidenzintervall-Band, das sich bis 2018 deutlich verengt. Im selben Jahr weist das Modell auch den niedrigsten CRPS-Wert von etwa 1.0 auf. Danach steigt der Wert bis 2020 wieder auf etwa 2.0, begleitet von einer Zunahme der Intervallbreite. Die Modelle Cosmo-1E und Cosmo-2E zeigen über den gesamten Zeitraum zickzackförmige CRPS-Verläufe mit wechselnden hohen und tiefen Werten zwischen 1.2 und 2.4, während die Breite der Konfidenzintervalle weitgehend konstant bleibt. Die Icon-Modelle (Icon-1 und Icon-2) weisen während ihrer Einsatzjahre 2024–2025 nahezu konstante CRPS-Werte auf. Lediglich bei Icon-2 ist im Jahr 2025 eine leichte Verbreiterung des Konfidenzintervalls erkennbar.

Bei einer Vorlaufszeit von 120 h zeigt Cosmo-LEPS zu Beginn ebenfalls den höchsten CRPS-Wert mit etwa 4.3, gefolgt von einem zickzackförmigen Rückgang bis 2015 auf rund 1.8, bevor der Wert im Jahr 2016 wieder auf etwa 3.4 ansteigt. Das Konfidenzintervall nimmt zunächst ab, wird im letzten Jahr jedoch wieder breiter. Cosmo-E zeigt einen ähnlichen Verlauf mit einem Minimum um 2018 bei etwa 1.4, gefolgt von einem erneuten Anstieg. Cosmo-2E beginnt im Jahr 2020 mit einem tiefen CRPS-Wert und zeigt einen zickzackförmigen Anstieg bis 2024 auf rund 3.3, es weist dabei die engsten Konfidenzintervalle aller Modelle auf. Icon-2 startet bei etwa 2.4 und steigt bis 2025 auf 3, wobei das Konfidenzintervall ebenfalls leicht zunimmt.

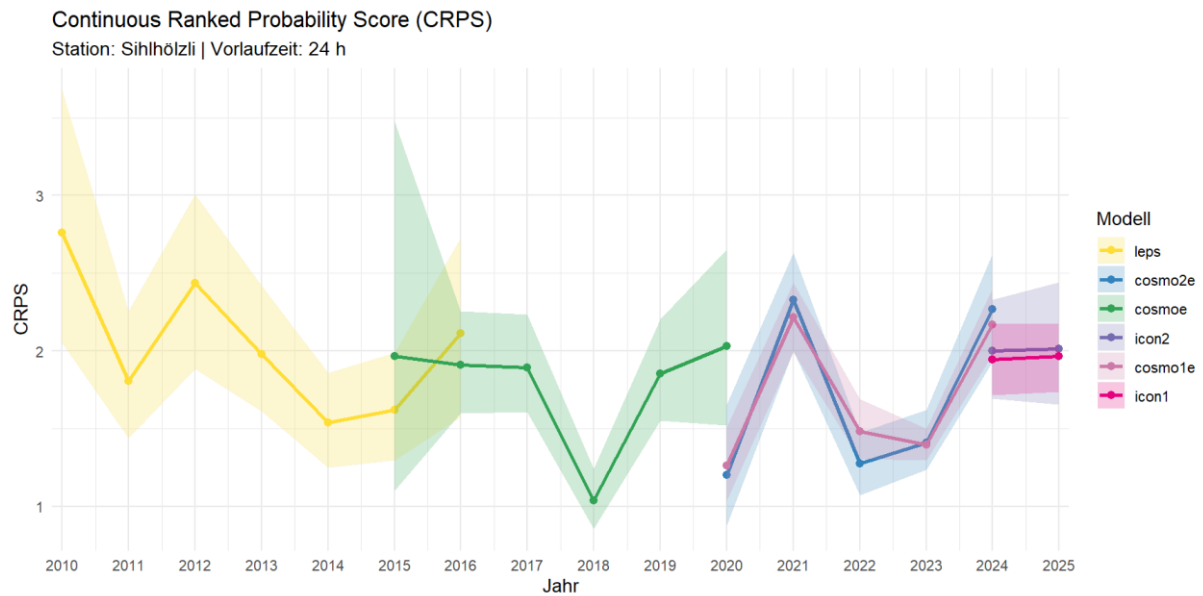


Abbildung 25: Zeitlicher Verlauf des mittleren Continuous Ranked Probability Score (CRPS) aller meteorologischen Modelle bei einer Vorlaufzeit von 24 h für die Station Sihlhölzli im Zeitraum Mai bis November 2010-2025. Die farbigen Linien zeigen die mittleren Jahreswerte pro Modell, die schattierten Bänder kennzeichnen die zugehörigen 90 %-Bootstrap-Konfidenzintervalle.

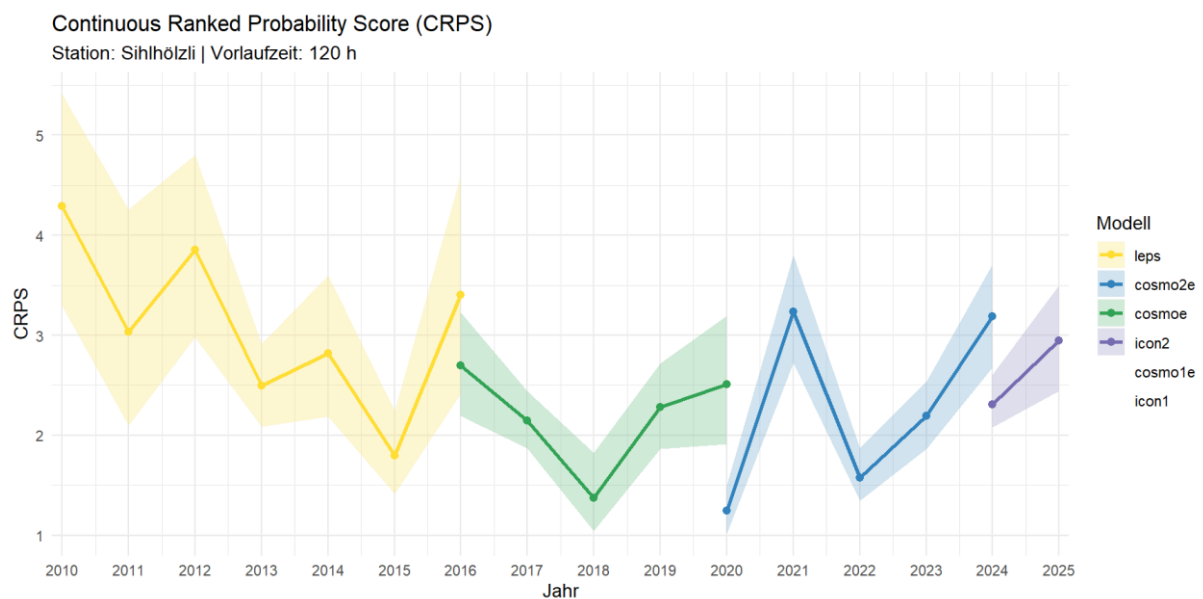


Abbildung 26: Zeitlicher Verlauf des mittleren Continuous Ranked Probability Score (CRPS) aller meteorologischen Modelle bei einer Vorlaufzeit von 120 h für die Station Sihlhölzli im Zeitraum Mai bis November 2010-2025. Die farbigen Linien zeigen die mittleren Jahreswerte pro Modell, die schattierten Bänder kennzeichnen die zugehörigen 90 %-Bootstrap-Konfidenzintervalle.

4.5.3 Probabilistischer Skill relativ zur Klimatologie

In Abbildung 27 und Abbildung 28 ist der zeitliche Verlauf des Continuous Ranked Probability Skill Score (CRPSS) der meteorologischen Modelle Icon-1, Icon-2, Cosmo-1E und Cosmo-2E dargestellt. Die schattierten Bänder zeigen die zugehörigen 90 %-Konfidenzintervalle und verdeutlichen die Unsicherheit der jeweiligen Mittelwerte. Positive Werte deuten auf eine bessere probabilistische Güte im Vergleich zur Klimareferenz hin, negative Werte auf eine schlechtere Leistung. Für die Icon-Modelle (Abbildung 27) beginnen die CRPSS-Werte bei etwa 0.25 und nehmen mit zunehmender Vorlaufzeit ab. Icon-1 sinkt bis rund 0.13 bei 32 h, während Icon-2 bei etwa 66 h die Nulllinie unterschreitet. Danach steigt der Wert leicht an, fällt jedoch bei etwa 114 h erneut unter die Nulllinie. Das Konfidenzintervall von Icon-2 bleibt ab etwa 18 h über die gesamte Vorlaufzeit hinweg weitgehend konstant, während das Intervall von Icon-1 insgesamt schmaler ausfällt.

Für die Cosmo-Modelle (Abbildung 28) beginnen die CRPSS-Werte sehr niedrig bei etwa -0.35. Dieser Wert ist vermutlich durch einen Ausreisser geprägt, da der Score im nächsten Zeitschritt bereits auf etwa -0.1 ansteigt. Anschliessend erhöhen sich die CRPSS-Werte beider Modelle bis zur Nulllinie bei rund 13 h, bevor sie wieder abnehmen. Ab etwa 36 h zeigt Cosmo-2E einen weitgehend konstanten Verlauf mit kleineren zickzackförmigen Schwankungen, jedoch ohne klaren Trend. Lediglich beim vorletzten Zeitschritt springt der Wert von rund -0.25 auf -0.06, was ähnlich wie beim ersten Zeitschritt auf einen Ausreisser hindeutet. Beide Cosmo-Modelle verlaufen über den grössten Teil der gemeinsamen Laufzeit nahezu deckungsgleich. Die Konfidenzintervalle sind zu Beginn bei beiden Modellen ähnlich breit, ab etwa 24 h nimmt die Breite bei Cosmo-2E leicht zu, bleibt danach jedoch konstant bis zum Ende der Vorhersageperiode.

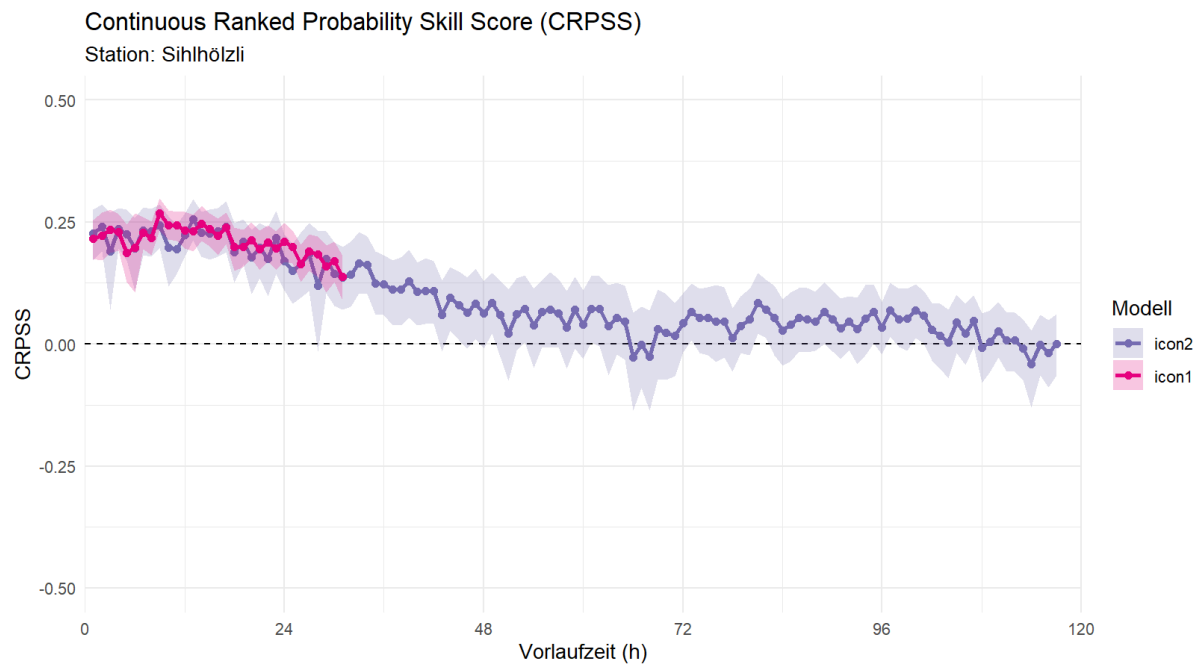


Abbildung 27: Zeitlicher Verlauf des Continuous Ranked Probability Skill Score (CRPSS) der meteorologischen Modelle Icon-1 und Icon-2 für die Station Sihlhölzli im Zeitraum Mai bis November 2024-2025. Die farbigen Linien zeigen die mittleren CRPSS-Werte pro Vorlaufszeit, die schattierten Bänder kennzeichnen die zugehörigen 90 %-Bootstrap-Konfidenzintervalle. Die gestrichelte Nulllinie markiert die Referenz zur Klimatologie.

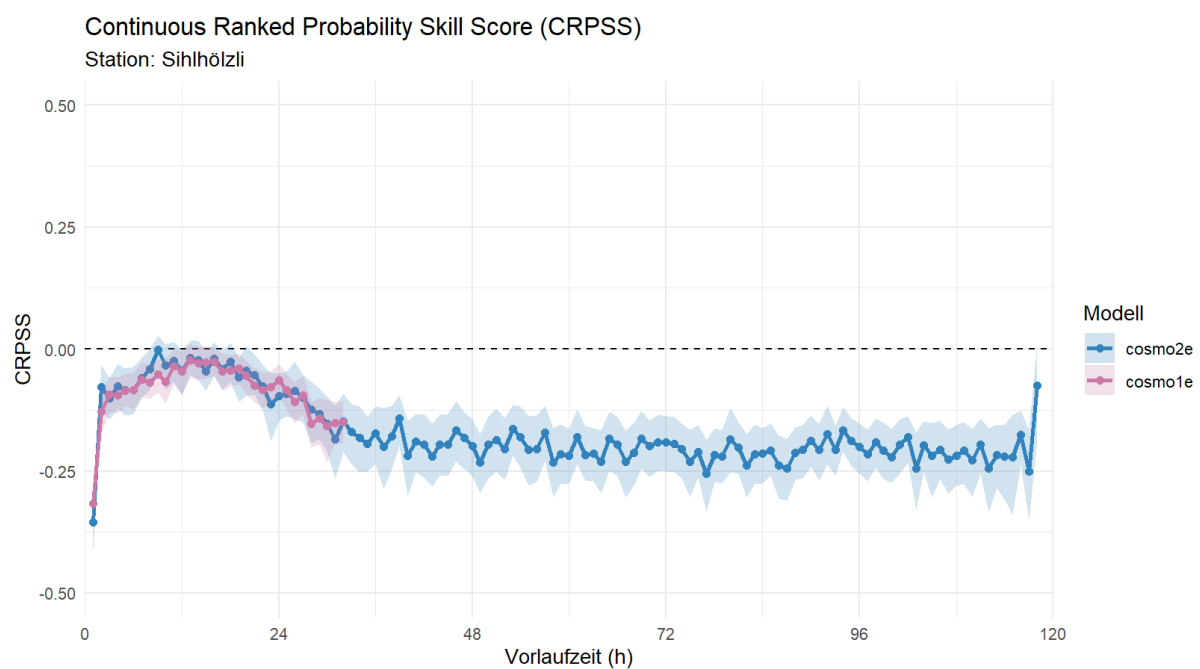


Abbildung 28: Zeitlicher Verlauf des Continuous Ranked Probability Skill Score (CRPSS) der meteorologischen Modelle Cosmo-1E und Cosmo-2E für die Station Sihlhölzli im Zeitraum Mai bis November 2020-2024. Dargestellt sind die mittleren CRPSS-Werte pro Vorlaufszeit mit den zugehörigen 90 %-Bootstrap-Konfidenzintervallen. Die gestrichelte Nulllinie zeigt die Referenz zur Klimatologie.

4.5.4 Kalibrierung der Ensembleverteilungen

In Abbildung 29 und Abbildung 30 sind die Rang-Histogramme der meteorologischen Modelle Icon-2 und Cosmo-2E für verschiedene Vorlaufzeiten (24 h, 48 h, 72 h, 118 h) dargestellt. Für Icon-2 (Abbildung 29) liegt die Fallzahl bei $N = 931$. Die höchsten Häufigkeiten treten bei den äussersten Rängen (Rang 0 und Rang 21) auf, mit relativen Häufigkeiten von jeweils etwa 0.4, während die mittleren Ränge nahezu keine Fälle aufweisen. Ab 48 h weist Rang 0 über alle Vorlaufzeiten hinweg den deutlich höchsten Frequency-Wert auf, wobei sich diese Tendenz mit Vorlaufzeit weiter verstärkt. Bei 118 h sinkt die Fallzahl leicht auf $N = 895$, das Verteilungsmuster bleibt jedoch ähnlich, mit einer ausgeprägten Häufung an den Randklassen.

Für Cosmo-2E (Abbildung 30) beträgt die Fallzahl über die meisten Vorlaufzeiten hinweg $N = 3245$, lediglich bei 119 h liegt sie bei $N = 1292$. Auch hier zeigen die äussersten Ränge die höchsten Häufigkeiten. Besonders Rang 0 weist in allen Vorlaufzeiten den deutlich höchsten Frequency-Wert von rund 0.6 auf, während Rang 21 eine deutlich geringere Häufigkeit von etwa 0.1-0.2 erreicht. Die mittleren Ränge weisen nur geringe Anteile auf. Das generelle Muster mit hohen Frequenzen an den Verteilungsrändern bleibt über alle Vorlaufzeiten hinweg bestehen.

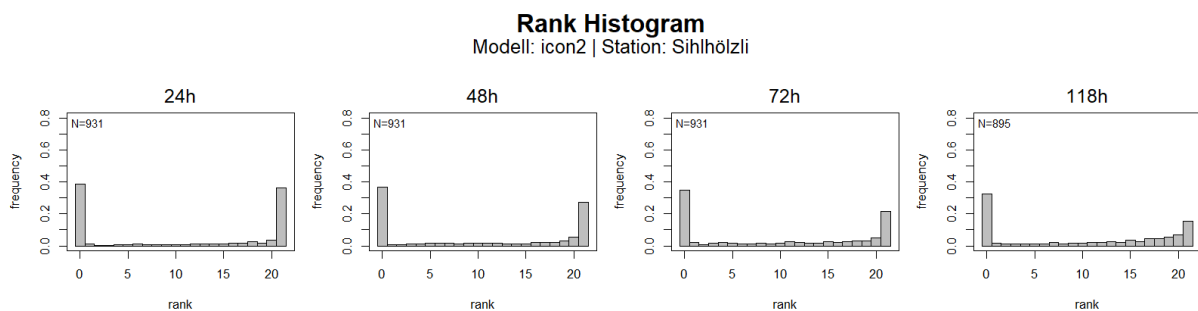


Abbildung 29: Rang-Histogramme des meteorologischen Modells Icon-2 für verschiedene Vorlaufzeiten (24 h, 48 h, 72 h und 118 h) an der Station Sihlhölzli im Zeitraum Mai bis November 2024–2025. Dargestellt ist die relative Häufigkeit der Ränge der Beobachtungen (N) innerhalb der Ensembleverteilung.

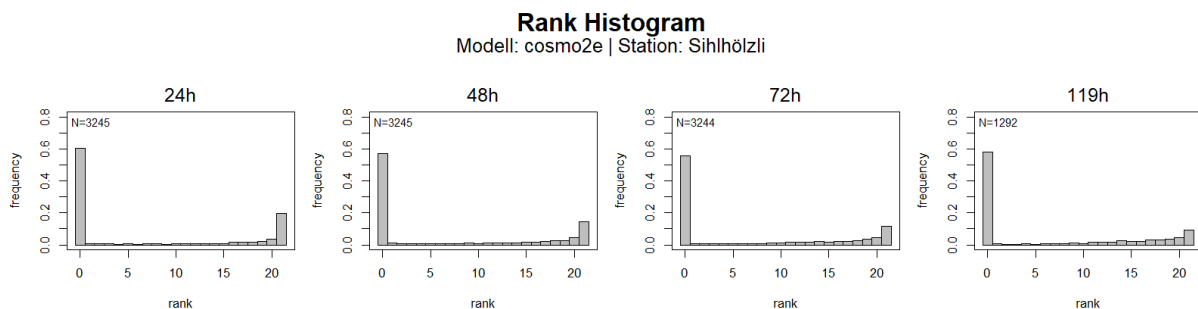


Abbildung 30: Rang-Histogramme des meteorologischen Modells Cosmo-2E für verschiedene Vorlaufzeiten (24 h, 48 h, 72 h und 118 h) an der Station Sihlhölzli im Zeitraum Mai bis November. Dargestellt ist die relative Häufigkeit der Ränge der Beobachtungen innerhalb der Ensembleverteilung.

4.6 Auswertungen für den Sihlsee

Die probabilistische Auswertung des Sihlseepegels zeigt im Vergleich zur dynamischen Abflussentwicklung am Sihlhölzli ein deutlich stabileres Vorhersageverhalten. In Abbildung 31 und Abbildung 32 ist der zeitliche Verlauf des Continuous Ranked Probability Skill Score (CRPSS) für die Modelle Icon-1, Icon-2, Cosmo-1E und Cosmo-2E dargestellt. Alle Modelle erreichen durchgehend positive CRPSS-Werte, was auf eine höhere probabilistische Güte im Vergleich zur Klimatologie hinweist. Die Icon-Modelle (Abbildung 31) zeigen zu Beginn sehr hohe CRPSS-Werte (≈ 0.9), die mit zunehmender Vorlaufszeit moderat abnehmen. Bei Icon-1 bleibt der Skill auch nach 32 h (≈ 0.75) stabil, während Icon-2 mit längerer Laufzeit eine stärkere Abnahme bis ≈ 0.1 aufweist. Die Unsicherheiten bleiben insgesamt gering, was auf eine konsistente Modellleistung über den Vorhersagehorizont hinweist.

Für die Cosmo-Modelle (Abbildung 32) beginnen die CRPSS-Werte sehr hoch, fast bei 1.0. Anschliessend nehmen die CRPSS-Werte beider Modelle ab, aber deutlich flacher als bei Icon-1 und Icon-2. Ab etwa 72 h zeigt Cosmo-2E einen sehr engen Konfidenzintervall, der sehr leicht zunimmt aber deutlich weniger breit als bei Icon-2 ist. Lediglich zwischen dem vorletzten und letzten Zeitschritt springt der CRPSS-Wert von Cosmo-2E von rund 0.38 auf 0.55, was auf einen Ausreisser hindeutet. Beide Cosmo-Modelle verlaufen über den grössten Teil der gemeinsamen Laufzeit nahezu deckungsgleich.

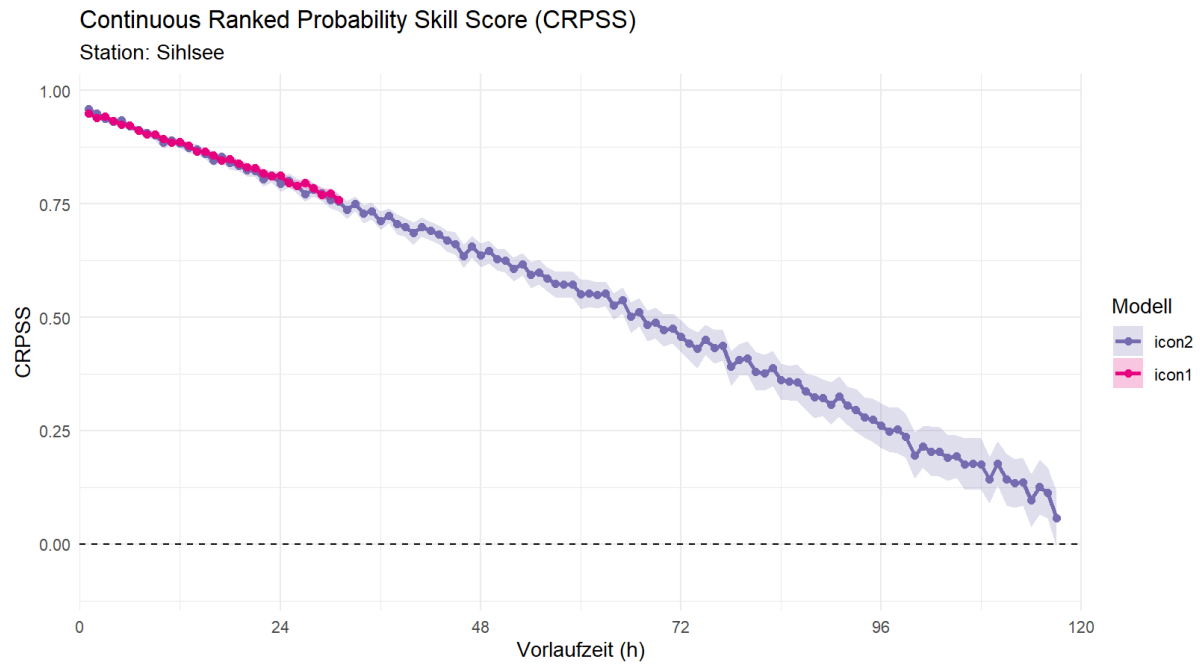


Abbildung 31: Zeitlicher Verlauf des Continuous Ranked Probability Skill Score (CRPSS) der meteorologischen Modelle Icon-1 und Icon-2 für die Station Sihlsee im Zeitraum Mai bis November 2024-2025. Die farbigen Linien zeigen die mittleren CRPSS-Werte pro Vorlaufzeit, die schattierten Bänder kennzeichnen die zugehörigen 90 %-Bootstrap-Konfidenzintervalle. Die gestrichelte Nulllinie markiert die Referenz zur Klimatologie.

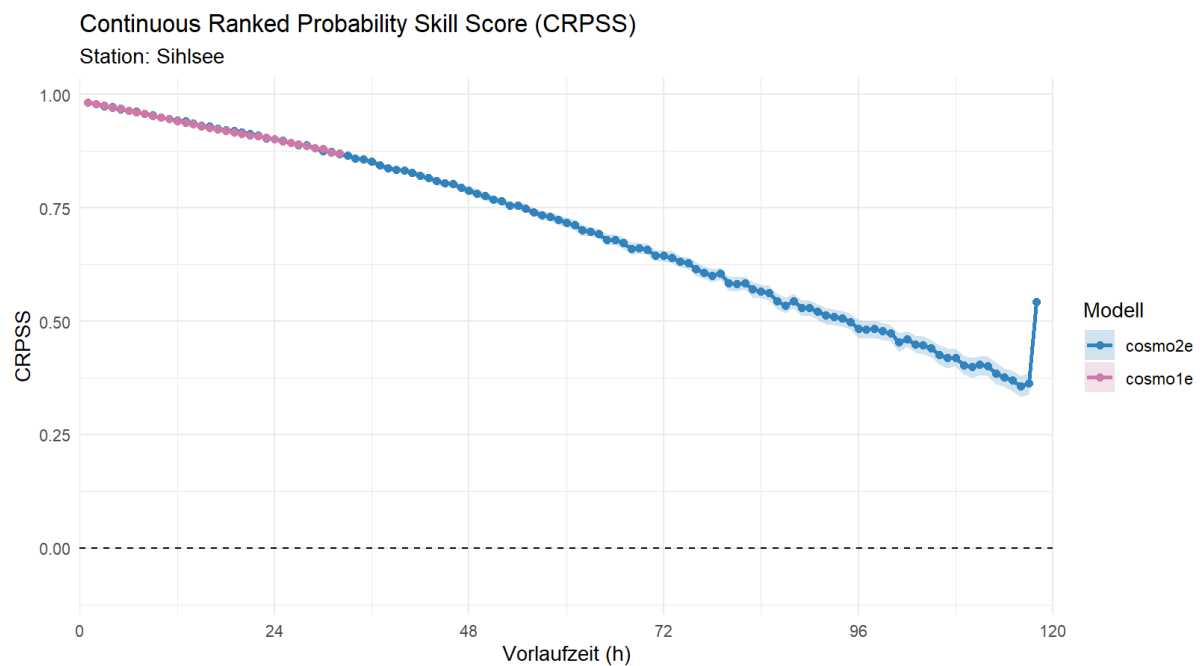


Abbildung 32: Zeitlicher Verlauf des Continuous Ranked Probability Skill Score (CRPSS) der meteorologischen Modelle Cosmo-1E und Cosmo-2E für die Station Sihlsee im Zeitraum Mai bis November 2020-2024. Die farbigen Linien zeigen die mittleren CRPSS-Werte pro Vorlaufzeit, die schattierten Bänder kennzeichnen die zugehörigen 90 %-Bootstrap-Konfidenzintervalle. Die gestrichelte Nulllinie markiert die Referenz zur Klimatologie.

5. Diskussion

5.1 Gesamtbewertung der Modellgüte

Die in Abbildung 8 und Abbildung 12 dargestellten Ergebnisse verdeutlichen, dass die deterministische Modellgüte der meteorologischen Vorhersagen für das Sihlgebiet in allen betrachteten Modellgenerationen ein konsistentes Grundmuster aufweisen. Mit zunehmender Vorlaufzeit nimmt die Vorhersagegüte erwartungsgemäss ab, wobei die Stärke dieses Rückgangs deutlich zwischen den Wettermodellen variiert. Die Analyse zeigt, dass kurze Vorlaufzeiten (24-48 h) durchwegs die höchsten Gütwerte erzielen, während längere Prognosen (≥ 72 h) eine deutliche Abnahme der Leistungsfähigkeit aufweisen. Die Entwicklung von 2010 bis 2025 spiegelt zugleich die schrittweise Ablösung älterer durch neue Modellgenerationen wider: Cosmo-LEPS, Cosmo-2 und Cosmo-7 waren von 2010 bis 2016 aktiv, Cosmo-1 und Cosmo-E von 2015 bis 2020, Cosmo-1E und Cosmo-2E von 2020 bis 2024 und schliesslich Icon-1 und Icon-2 ab 2024. Diese unterschiedlichen Aktivitätszeiträume beeinflussen die Vergleichbarkeit der statistischen Streuung, da die Icon-Modelle zum Zeitpunkt dieser Studie nur rund 1.5 Jahre abdecken und somit auf einer deutlich kleineren Datengrundlage beruhen.

In Abbildung 8, welche die NSE über die Jahre und Vorlaufzeiten zeigt, wird ein deutlicher Rückgang der Modellgüte mit wachsender Prognosedauer sichtbar. Die frühen Cosmo-Wettermodelle zeigen eine stärkere Abnahme der NSE-Werte bereits ab 48 Stunden, während Cosmo-LEPS die höchsten Medianwerte und eine geringe Streuung aufweist. Diese geringe Variabilität deutet auf eine stabile, jedoch leicht gedämpfte Reaktion auf extreme Abflussereignisse hin, was bereits in früheren Analysen von Addor et al. (2011) und Liechti und Zappa (2012, 2014) als typische Eigenschaft des Cosmo-LEPS beschrieben wurde. Die neueren Systeme Cosmo-E und Cosmo-2E zeigen dagegen eine breitere Streuung der Jahres-NSE, was auf eine grössere Variabilität in der Güte über die Jahre hinweist. Icon-2 zeigt ab einer Vorlaufzeit von 72 Stunden eine engere Verteilung der NSE-Werte auf. Aufgrund der kurzen Aktivitätsdauer ist diese enge Streuung jedoch nur bedingt vergleichbar mit den länger betriebenen Cosmo-Modellen. Sie deutet zwar auf eine konsistente Modellleistung in den verfügbaren Jahren hin, lässt aber noch keine belastbare Aussage über die interannuelle Stabilität zu. Insgesamt bestätigt die beobachtete Abnahme der NSE-Mediane unter 0.5 ab 48 Stunden das bekannte Muster einer abnehmenden deterministischen Vorhersagegüte mit wachsender Vorlaufzeit, das bereits in früheren Evaluierungen der Sihl-Vorhersagesysteme dokumentiert wurde (Addor et al., 2011; Liechti & Zappa, 2012, 2014).

Die in Abbildung 9 dargestellten RMSE-Werte zeigen ein ähnliches Muster: Die Fehler nehmen mit der Vorlaufzeit zu, wobei die Streubreite zwischen den meteorologischen Modellen stark variiert. Icon-2 weist ab 72 Stunden die geringsten Medianwerte und die engste Verteilung auf. Bei kürzeren Vorlaufzeiten (24-48 h) erzielt Cosmo-2E die geringsten Medianwerte, während Cosmo-LEPS höhere Fehler, aber eine geringere Streuung zeigt. Liechti et al. (2016) beobachteten für die Übergangsphase zwischen Cosmo-LEPS und Cosmo-E in ähnliches Verhalten, bei dem die höhere räumliche Auflösung zwar die Medianfehler reduzierte, die Streuung jedoch zunahm, da konvektive Ereignisse stärker aufgelöst wurden. Dieses Muster bestätigt sich in den vorliegenden Ergebnissen.

Die systematischen Fehler in Abbildung 10 verdeutlichen eine tendenzielle Unterschätzung der Abflüsse über alle Wettermodelle hinweg. Die überwiegend negativen Medianwerte des ME zeigen, dass die Modelle den beobachteten Abfluss meist zu niedrig prognostizieren. Mit zunehmender Vorlaufzeit verstärkt sich diese negative Verzerrung leicht, was darauf hinweist, dass die Modelle mit wachsender Prognosedauer zunehmend konservativer werden. Dieses Verhalten wurde bereits in den Studien von Addor et al. (2011) und Liechti und Zappa (2014) beschrieben und auf die begrenzte

Fähigkeit der damaligen Cosmo-Konfigurationen zur Reproduktion intensiver konvektiver Niederschläge zurückgeführt.

Die linearen Zusammenhänge zwischen Beobachtungen und Vorhersagen (Abbildung 11, Abbildung 12 und Tabelle 3) bestätigen diese Muster. Für kurze Vorlaufzeiten (Tag 1) liegen die R^2 -Werte über alle Modellgenerationen hinweg im Bereich 0.7 bis 0.9. Bei längeren Prognosen (Tag 5) sinken die R^2 -Werte deutlich, bewegen sich in den neueren Modellen aber auf stabilerem Niveau. Während Cosmo-LEPS R^2 -Werte zwischen 0.25-0.58 aufweist, erreicht Cosmo-2E R^2 -Werte zwischen 0.12 und 0.67 und weist damit eine ausgeprägte interannuelle Variabilität auf. Für Icon-2 liegen die Werte im Bereich von 0.44 bis 0.52 mit engerer Streuung. Diese scheinbar höhere Stabilität von Icon-2 beruht jedoch auf der kurzen Aktivitätsdauer und ist daher nicht direkt mit den länger betriebenen Cosmo-Systemen vergleichbar. Liechti und Zappa (2020) stellten für den Zeitraum 2015-2019 eine weitgehend stabile, aber nicht weiter zunehmende Modellgüte fest, was mit den hier beobachteten Entwicklungen übereinstimmt.

Zusammenfassend zeigt die Analyse der deterministischen, kontinuierlichen Gütemassen, dass sich die kurzfristige Vorhersagegüte der meteorologischen Modelle im Sihlgebiet im Verlauf der letzten 15 Jahre auf hohem Niveau stabilisiert hat, während sich die längerfristigen Prognosen durch geringere Streuung und robustere Fehlerstrukturen verbessert haben. Diese Entwicklung muss jedoch im Lichte der unterschiedlichen Modellaktivitätszeiträume betrachtet werden: Die engere Streuung der Icon-Modelle beruht auf einer deutlich kürzeren Beobachtungsperiode und kann daher nicht direkt mit den länger betriebenen Cosmo-Systemen verglichen werden. Fortschritte in der numerischen Wettervorhersage äussern sich somit weniger in einer allgemeinen Erhöhung der Gütewerte als in einer verbesserten Konsistenz und Verlässlichkeit der Vorhersagen, was im Einklang mit den Ergebnissen von Addor et al. (2011) und Liechti und Zappa (2020) steht.

5.2 Ereignisbasierte Vorhersage und Schwellenanalyse

Die ereignisbasierte Bewertung der meteorologischen Vorhersagen zeigt, dass die Fähigkeit der Modelle, Abflussschwellen zuverlässig zu erkennen, stark von der Schwellenhöhe und der Prognosedauer abhängt. Sowohl der HSS als auch der CSI nehmen mit steigender Schwelle und längerer Vorlaufzeit deutlich ab (Abbildung 13 und Abbildung 14). Die absolute Höhe von HSS und CSI muss im Kontext ihrer Skala interpretiert werden. Beide Metriken reichen für die betrachtete Studienzeit von 0 bis 1, wobei ein Wert von 1 eine perfekte Übereinstimmung zwischen Vorhersage und Beobachtung bedeutet. Werte nahe 0 zeigen hingegen, dass die Trefferquote kaum besser ist als ein zufälliger Entscheidungsprozess. Somit gilt: Je näher die Werte an 1 liegen, desto höher ist die Fähigkeit des Modells, Ereignisse korrekt zu erkennen und gleichzeitig Fehlalarme zu vermeiden. Ein niedriger, aber positiver Wert deutet nicht zwingend auf eine «gute» Prognose hin, sondern kann auch eine schwache, kaum nutzbare Erkennungsleistung widerspiegeln. Die Abnahme von HSS und CSI mit zunehmender Schwellenhöhe verdeutlicht die strukturelle Herausforderung der meteorologischen Modelle, Extremereignisse realistisch zu erfassen. Insbesondere die kritische operative Schwelle von 40 m³/s am Limmatwehr wird nur unvollständig erkannt. Während bei häufigeren, niedrigeren Abflüssen (10 und 20 m³/s) noch moderate Übereinstimmungen zwischen Beobachtungen und Prognosen bestehen, bricht die Vorhersagegüte bei höheren Schwellen deutlich ein. Diese Entwicklung deutet darauf hin, dass die Modelle zwar in der Lage sind, den allgemeinen Verlauf hydrologischer Ereignisse zu reproduzieren, aber Schwierigkeiten haben, die Dynamik seltener Spitzenabflüsse präzise abzubilden. Das meteorologische Modell Icon-2 zeigt tendenziell stabilere Streuungen der Gütemasse, jedoch nicht zwingend höhere Werte. Die geringere Variabilität beruht jedoch auf einer verkürzten

Beobachtungsperiode und kann nicht als belastbarer Hinweis auf eine erhöhte Robustheit gewertet werden.

Die persistierende Schwellenabhängigkeit der Modellgüte wurde bereits in den früheren Auswertungen des Sihl-Vorhersagesystems festgestellt. Addor et al. (2011) zeigten, dass die Zuverlässigkeit der Cosmo-basierten Vorhersagen insbesondere bei hohen Abflüssen begrenzt ist. Die Autoren führen dies auf die starke Abhängigkeit des hydrologischen Modells von der Güte der Niederschlagsvorhersage zurück, wobei Fehler in Intensität und räumlicher Verteilung der Niederschläge in komplexer Topographie zu einer systematischen Unterschätzung der Abflussspitzen führen. Liechti und Zappa (2014) bestätigten dieses Muster und führten es auf die Limitierungen der damaligen modellbasierten Niederschlagsvorhersage im regionalen Massstab zurück. Trotz Fortschritten in der physikalischen Modellierung und der Ensemblemethodik bleibt dieser Effekt auch im jüngsten Zeitraum erkennbar.

5.3 Probabilistische Bewertung

Die ROC-Kurven (Abbildung 17 und Abbildung 18) zeigen für niedrige Schwellen (10-20 m³/s) eine ausgeprägt konvexe Form und damit eine hohe Trennschärfe zwischen Ereignis und Nicht-Ereignis. Mit zunehmender Vorlaufzeit flachen die ROC-Kurven ab, und bei höheren Schwellen (≥ 40 m³/s) werden sie stufig. Dies ist ein typisches Zeichen seltener Datenlagen, denn wenige Ereignisse führen zu diskreten Sprüngen der TPR/FPR und damit zu statistisch instabilen Kurvensegmenten. Dieser Befund spiegelt sich in den AUC-Werten bei 24 h (Abbildung 19) wider, die für alle Wettermodelle deutlich oberhalb des Zufallsniveaus liegen (> 0.5). Im Langfristbereich (118 h, Abbildung 20) divergiert die AUC-Entwicklung der Modellgenerationen, die Trennschärfe nimmt ab und die interannuelle Variabilität nimmt zu. Während einige Wettermodelle AUC-Werte nahe 1 ansteigen, zeigen andere zeitweise Einbrüche bis in die Nähe der Zufallsgrenze (Cosmo-E um 2020). Über die Jahre zeigen ältere Cosmo-Generationen teils grössere AUC-Schwankungen, während Cosmo-2E und Icon-2 tendenziell stabilere (aber nicht durchgängig höhere) AUC-Niveaus erreichen. Dies bedeutet, dass Ereignisse kurzfristig gut trennbar sind, die Trennschärfe für seltene oder hohe Schwellen mit Vorlaufzeit deutlich abnimmt. Diese Muster sind konsistent mit der Erwartung, dass ROC und AUC primär die Diskrimination erfassen, jedoch nicht die Kalibrierung (Over- oder Underconfidence) bewerten. So kann ein System also hohe AUC-Werte aufweisen und dennoch schlecht kalibriert sein. Genau in diesem Sinn wird ROC und AUC in der Verifikationspraxis als potenzieller Skill interpretiert, der erst durch geeignete Kalibrierung in nutzbare Wahrscheinlichkeiten übersetzt wird (Brown et al., 2010).

Die Attributdiagramme (Abbildung 21 und Abbildung 22) erlauben eine detaillierte Bewertung der Kalibrierung und Zuverlässigkeit der Ensemblevorhersagen und bilden damit einen zentralen Bestandteil der probabilistischen Modellbewertung. Sie zeigen, wie gut die vorhergesagten Wahrscheinlichkeiten mit den tatsächlich beobachteten Eintrittshäufigkeiten übereinstimmen und ob systematische Tendenzen zu Über- oder Unterschätzung bestehen. Für operationelle Nutzer wie das AWEL sind diese Darstellungen besonders relevant, da sie unmittelbar die Glaubwürdigkeit und Nutzbarkeit der Wahrscheinlichkeitsprognosen widerspiegeln. Beim Modell Icon-2E liegen die Punkte der ersten vier Vorhersagegruppen (1-24 h, 25-48 h, 49-72 h, 73-96 h) überwiegend unterhalb der 1:1 Diagonalen, was auf eine systematische Überschätzung der Eintrittswahrscheinlichkeit (Overforecasting) hinweist. Das heisst, Ereignisse werden mit höheren Wahrscheinlichkeiten prognostiziert, als sie tatsächlich eintreten. Erst in der längsten Vorhersagegruppe (97-120 h) kehrt sich dieses Muster um. Die Punkte liegen teilweise oberhalb der Diagonalen, insbesondere bei hohen p-Werten ($p > 0.7$). Dies weist auf eine zunehmende Underconfidence bei längeren Vorlaufzeiten hin, das heisst auf eine Unterschätzung der Eintrittswahrscheinlichkeit seltener Ereignisse. Mit zunehmender

Vorlaufzeit und höherer Schwelle nehmen auch die Konfidenzintervalle der einzelnen Wahrscheinlichkeitsklassen deutlich zu. Diese Streuung reflektiert eine wachsende statistische Unsicherheit infolge des begrenzten Stichprobenumfangs. Wie Liechti und Zappa (2020) hervorheben, erfordert die Unterteilung der Prognosen in Wahrscheinlichkeitsklassen eine grosse Datenmenge, um stabile Aussagen zur Kalibrierung zu ermöglichen. Besonders bei seltenen Hochwasserereignissen wird die Analyse unsicher, da nur wenige Fälle pro Klasse auftreten was, wie bereits in Kapitel 5.1 erläutert, auch durch den kürzeren Beobachtungszeitraum beeinflusst sein dürfte. Im operationellen Kontext bedeutet das, dass obwohl Icon-2E eine erkennbare positive Beziehung zwischen vorhergesagter und beobachteter Wahrscheinlichkeit zeigt (steigende Kurvenpunkte), deuten die Abweichungen unterhalb der Diagonalen bei kurzen Vorlaufzeiten auf eine zu hohe Warnwahrscheinlichkeit hin. Für den Endnutzer kann dies den Eindruck einer überhöhten Hochwassergefahr erwecken. Die Modellunsicherheit ist insbesondere bei längeren Vorhersagen (ab 72 h) stark p-abhängig und sollte bei der operativen Entscheidungsfindung explizit berücksichtigt werden. Brown et al. (2010) betonen, dass eine Kombination aus Over- und Underconfidence oft aus unterdispersen Ensembles resultiert, deren Mitglieder die Beobachtung zu selten überdecken. Ein Befund, der sich in den Rang-Histogrammen (U-Form in den Abbildung 29 und Abbildung 30) klar bestätigt. Das Modell Cosmo-2E zeigt eine über weite Bereiche nahezu ideale Kalibrierung. Die Punkte der ersten beiden Vorhersagegruppen (1-24 h, 25-48 h) liegen eng an der 1:1 Diagonalen, was auf eine realistische Abbildung der Eintrittswahrscheinlichkeiten hinweist. Ab etwa 25-48 h zeigt sich jedoch ein S-förmiges Muster: Punkte mittlerer Wahrscheinlichkeiten ($p \approx 0.25-0.6$) liegen oberhalb der Diagonalen (Underforecasting), während hohe Wahrscheinlichkeiten ($p > 0.6$) unterhalb der Diagonalen liegen (Overforecasting). Dieses Verhalten entspricht dem in der Literatur beschriebenen bedingten Bias, bei dem Modelle in mittleren Bereichen zu konservativ und in hohen Bereichen zu optimistisch sind (Brown et al., 2010; Wilks, 2006). Ein wesentlicher Grund dafür ist die ungleiche Besetzung der Wahrscheinlichkeitsklassen. Der grösste Teil der Vorhersagen fällt in die niedrigste Klasse ($p < 0.2$), während höhere Klassen nur sehr spärlich besetzt sind. Bereits Liechti und Zappa (2020) wiesen darauf hin, dass diese ungleiche Datenverteilung die statistische Unsicherheit stark erhöht und eine präzise Beurteilung der Kalibrierung erschwert. In den Abbildungen äussert sich dies in den breiten Konfidenzintervallen der höheren p-Bereiche, die auf die geringe Fallzahl zurückzuführen sind. Dennoch bleibt die Beziehung zwischen Vorhersagewahrscheinlichkeit und beobachteter Häufigkeit monoton steigend, und die meisten Punkte liegen innerhalb des grau markierten Toleranzbereichs, der laut Liechti und Zappa (2020) als praktisch nutzbar gilt. Damit erfüllt Cosmo-2E die Anforderungen an ein operativ brauchbares Vorhersagesystem, insbesondere im kurzfristigen Bereich bis 48 h.

Mit dem BSS (Abbildung 15 und Abbildung 16) lässt sich die probabilistische Gesamtgüte der Ensembleprognosen beurteilen. In der vorliegenden Analyse wurde der Ensemble-Median als deterministische Prognose behandelt und der BSS relativ zur Klimatologie berechnet. Damit misst der BSS, ob die Medianvorhersage eine Verbesserung gegenüber einer rein klimatologischen Schätzung darstellt (Wilks, 2006). Wie zu erwarten, nehmen die BSS-Werte mit zunehmender Vorlaufzeit ab, was die wachsende Unsicherheit in den Ensembleprognosen widerspiegelt. Dennoch bleiben die Werte der Cosmo-Modelle bis zu einer Schwelle von $40 \text{ m}^3/\text{s}$ über den gesamten Fünf-Tage-Zeitraum hinweg positiv, was bedeutet, dass das System auch am letzten Vorhersagetag noch einen probabilistischen Mehrwert gegenüber der Klimatologie bietet. Bei $60 \text{ m}^3/\text{s}$ fällt der Cosmo-2E Median in den Bereich um die Nulllinie und bei $80 \text{ m}^3/\text{s}$ erreichen die Cosmo-2E BSS-Werte die Nulllinie, das heisst der probabilistische Nutzen gegenüber einer klimatologischen Schätzung verschwindet bei sehr hohen Abflüssen. Insgesamt liegen Cosmo-1E und Cosmo-2E im Bereich der Schwellen $10-40 \text{ m}^3/\text{s}$ auf vergleichbarem Niveau. Der praktische Vorteil des neueren Systems liegt daher primär in der höheren Aktualisierungsfrequenz und verbesserten räumlichen Auflösung (Liechti & Zappa, 2020). Im Gegensatz dazu zeigen die Icon-Modelle ein weniger konsistentes Bild: Während die Ensemble- und Median-BSS-

Werte für 10 m³/s zunächst noch positiv bleiben, fallen sie bei 20 m³/s mit zunehmender Vorlaufzeit schnell ab und der Icon-2 Median erreicht bereits nach rund 30 h negative Werte. Für operationelle Anwender wie das AWEL hat der rasche Rückgang der BSS-Werte der Icon-Modelle unmittelbare Konsequenzen. Während Cosmo-2E auch im mehrtägigen Vorhersagebereich (bis fünf Tage) durchgehend positive BSS-Werte zeigt und damit gegenüber der Klimatologie einen messbaren Mehrwert bietet, verlieren die Icon-Modelle diesen Vorteil bereits sehr früh. Besonders bei der für das Hochwassermanagement entscheidenden Schwelle von 40 m³/s erreichen die Icon-BSS-Werte schon nach etwa 18 Stunden die Nulllinie und fallen danach in den negativen Bereich. Das bedeutet, dass die Vorhersagen von Icon-1 und Icon-2 ab diesem Zeitpunkt keine Verbesserung mehr gegenüber einer klimatologischen Schätzung darstellen, sondern probabilistisch sogar schlechter abschneiden. Der rasche Rückgang der BSS-Werte der Icon-Modelle weist auf Unterdispersion hin, also eine zu geringe Ensemble-Streuung relativ zur tatsächlichen Variabilität und diese unzureichende Streuung führt dazu, dass Wahrscheinlichkeiten systematisch überschätzt und Extremereignisse zu wenig differenziert werden (Brown et al., 2010).

Diese Unterdispersion ist in den Rang-Histogrammen (Abbildung 29 und Abbildung 30) deutlich sichtbar, wo die Beobachtungen überdurchschnittlich häufig an den Verteilungsrändern liegen (U-Form). Dies bedeutet, dass das beobachtete Abflussverhalten häufiger ausserhalb des Prognosebereichs liegt, als es bei einem perfekt kalibrierten probabilistischen Wettermodell zu erwarten wäre (Liechti & Zappa, 2014). Bei Cosmo-2E ist zusätzlich eine leichte Asymmetrie sichtbar. Die Beobachtungen liegen häufiger unterhalb als oberhalb des Ensembles, was auf einen leichten negativen Bias hinweist. Dieses Verhalten korrespondiert mit den Attributdiagrammen beobachteten Mustern. Icon-2 zeigt dagegen eine symmetrischere U-Form, die primär auf eine zu geringe Streuung hinweist. Beide Modelle erfassen somit den mittleren Abflussbereich recht konsistent, sind aber in den Extremen nicht ausreichend kalibriert. In Verbindung mit den Attributdiagrammen (Abbildung 21 und Abbildung 22) erklärt die Unterdispersion sowie die negativen BSS-Werte der Icon-Modelle die bei kurzen Vorlaufzeiten beobachtete Überschätzung der Eintrittswahrscheinlichkeit seltener Ereignisse. Für den Endnutzer bedeutet dies, dass Icon-basierte Wahrscheinlichkeitsprognosen ab etwa einem Tag Vorlaufzeit mit Vorsicht zu interpretieren sind und ohne zusätzliche statistische Nachkalibrierung keinen verlässlichen Mehrwert gegenüber der Klimatologie bieten (Brown et al., 2010).

Die kontinuierlichen probabilistischen Metriken ergänzen das Bild der Schwellenbewertung. Der CRPS (Abbildung 23-Abbildung 26) steigt bei allen Wettermodellen mit zunehmender Vorlaufzeit, was die wachsende Unsicherheit der Ensembleverteilungen widerspiegelt. In den CRPS-Verläufen (Abbildung 27 und Abbildung 28) zeigen sich Unterschiede: Die Icon-Modelle beginnen mit positiven Werten um 0.25, sinken aber rasch ab und erreichen nach rund 60 Stunden die Nulllinie. Die Cosmo-Modelle starten leicht negativ, steigen kurzzeitig auf 0 an und fallen ab etwa 20 Stunden wieder leicht ins Negative. Die Konfidenzintervalle sind bei Cosmo-2E minimal enger als bei Icon-2, der Unterschied bleibt jedoch gering. Dieses Muster lässt sich möglicherweise durch die Berechnung der Klimatologie, die als Referenz für den CRPS dient, erklären. Für jedes Modell wird eine eigene Klimatologie aus den Beobachtungen seines jeweiligen Zeitraums erstellt, bei Cosmo aus den Jahren 2020-2024, bei Icon lediglich aus 2024-2025. Dadurch basiert jede Bewertung auf einer modell- und zeitspezifischen Referenz, was die Vergleichbarkeit einschränken und Icon aufgrund der kürzeren, weniger stabilen Datengrundlage scheinbar bessere CRPS-Werte verleihen kann. Ideal wäre daher eine einheitliche Klimatologie, die auf einem längeren Zeitraum (zum Beispiel 2010-2025) beruht und für alle Modelle gleichermaßen gilt. Nur so liesse sich der CRPS direkt vergleichen und die tatsächliche Modellgüte verlässlich beurteilen. Für den operationellen Einsatz bedeutet dies, dass die kurzfristig höheren CRPS-Werte der Icon-Modelle nicht als eindeutiger Qualitätsvorteil interpretiert werden sollten, sondern vielmehr als Folge der unterschiedlichen Referenzgrundlage zu verstehen sind. Eine robuste

Beurteilung der praktischen Prognoseleistung sollte daher auf gemeinsame, konsistente Bewertungszeiträume und ergänzende Metriken gestützt werden, um verlässliche Aussagen über den tatsächlichen Nutzen im Vorhersagebetrieb zu ermöglichen.

Für den Sihlsee zeigen die meteorologischen Modelle insgesamt eine hohe und stabile probabilistische Vorhersagegüte. Abbildung 31 und Abbildung 32 veranschaulichen den zeitlichen Verlauf des CRPSS. Alle Wettermodelle erreichen im Kurzfristbereich (bis 24 h) hohe CRPSS-Werte zwischen 0.8 und 1.0, was auf eine sehr gute Übereinstimmung mit der Beobachtung und einen deutlichen Mehrwert gegenüber der Klimatologie hinweist (Brown et al., 2010). Da die klimatologische Referenz jedoch auf die Aktivitätsjahre der einzelnen Modelle beschränkt ist, sollten diese hohen Werte als relativer Skill innerhalb des jeweiligen Zeitraums interpretiert werden und nicht als absoluter Vergleich zwischen den Modellen. Mit zunehmender Vorlaufzeit nimmt der Skill nur moderat ab, was auf die geringe Variabilität und die starke Speicherwirkung des Sees zurückzuführen ist. Bei Icon-2 ist ab etwa 32 h ein leichter Anstieg der Unsicherheiten erkennbar, während die Cosmo-Modelle über nahezu die gesamte Laufzeit sehr enge Konfidenzintervalle aufweisen. Diese geringe Streuung spiegelt die hohe Stabilität und geringe meteorologische Sensitivität des regulierten Pegels wider. Damit bestätigt sich, dass der Sihlseepegel wesentlich träger und weniger dynamisch auf meteorologisch-hydrologische Veränderungen reagiert als der Abfluss bei der Messstation Sihlhölzli, gleichzeitig aber konstant gut vorhersagbar ist.

6. Ökonomische Bewertung und Entscheidungsrelevanz

Der ökonomische Wert eines Vorhersagesystems gibt an, in welchem Mass die erzeugten Vorhersagen die Entscheidungsfindung unterstützen und welchen Nutzen ein Anwender daraus ziehen kann (Liechti & Zappa, 2014). Eine direkte monetäre Quantifizierung des Nutzens wurde in dieser Arbeit nicht vorgenommen, doch aus den Ergebnissen lassen sich qualitative Schlussfolgerungen ableiten. Die BSS-Analyse zeigt, dass die früheren Cosmo-Modelle (insbesondere Cosmo-2E) bis zu einer Schwelle von 40 m³/s über den gesamten Fünf-Tage-Zeitraum hinweg positive Werte erreichten und somit einen klaren Mehrwert gegenüber der Klimatologie boten. Dies ermöglichte operationellen Akteuren wie dem AWEL, Entscheidungen über Speicherbewirtschaftung oder präventive Einsatzplanung auf einer belastbaren probabilistischen Grundlage zu treffen. Seit der Ablösung der Cosmo-Modellen durch die Icon-Modelle hat sich dieses Bild jedoch verändert. Die Icon-Modelle zeigen zwar kurzfristig eine hohe Schärfe der Prognosen, verlieren diesen Vorteil jedoch bereits nach 18 h, wenn die BSS-Werte in den neutralen oder negativen Bereich fallen. Ab diesem Zeitpunkt bieten die Vorhersagen keinen Mehrwert mehr gegenüber einer klimatologischen Einschätzung. Zudem weisen die Icon-Modelle eine geringere Kalibrierung und stärkere Unterdispersion auf, was zu systematischen Überschätzungen der Eintrittswahrscheinlichkeiten führt. Für den Hochwasserschutz bedeutet dies eine Verschlechterung der operationellen Zuverlässigkeit: Frühwarnungen können tendenziell zu häufig ausgelöst werden, während der probabilistische Informationsgehalt ab dem zweiten Vorhersagetag deutlich sinkt. Wenn die aktuelle Modellgeneration eine geringere Kalibrierung und Stabilität aufweist, steigt der Aufwand für die Interpretation der Prognosen und für ergänzende Bewertungen (zum Beispiel hydrologische Nachkalibrierung oder kombinierte Modellansätze) (Brown et al., 2010). Die Icon-Modelle bieten kurzfristig (bis etwa 24 h) weiterhin wertvolle Informationen, langfristig ist ihr ökonomischer Beitrag jedoch eingeschränkt und stark von der zukünftigen Integration statistischer Korrekturverfahren abhängig.

Besonders deutlich wird diese ökonomische Relevanz am Beispiel der Baustelle am Limmatwehr in Zürich, wo Entscheidungen des AWEL unmittelbar von den Abflussprognosen abhängen. Diese Einschätzungen beruhen auf einer Baustellenbegehung sowie einem fachlichen Austausch mit der zuständigen Person beim AWEL, die die operationellen Entscheidungsprozesse vor Ort erläuterte. Dabei wurde deutlich, dass fehlerhafte oder unzuverlässige Vorhersagen erhebliche finanzielle Konsequenzen haben können: Wird die Baustelle aufgrund einer prognostizierten Schwellenüberschreitung vorsorglich geschlossen, ohne dass der kritische Abfluss tatsächlich erreicht wird, entstehen vergleichbare wirtschaftliche Verluste wie bei einem tatsächlichen Hochwasserereignis, Stillstand des gesamten Baustellenbetriebs, Terminverzögerungen und zusätzliche Kosten. Nur im Fall, dass keine Schliessung erfolgt und die Schwelle tatsächlich nicht überschritten wird, entstehen keine direkten Kosten. Umgekehrt kann eine Unterschätzung der Hochwasserwahrscheinlichkeit dazu führen, dass bei tatsächlicher Schwellenüberschreitung materielle Schäden und Sicherheitsrisiken auftreten. Das AWEL steht somit vor einem permanenten Abwägungsdilemma zwischen wirtschaftlicher Effizienz und Risikominimierung. Je unzuverlässiger die probabilistische Vorhersage, wie derzeit bei den Icon-Modellen beobachtet, desto schwieriger und kostenintensiver wird diese Entscheidungsfindung.

7. Schlussfolgerungen und Ausblick

Über alle betrachteten Modellgenerationen zeigt sich ein konsistentes Grundmuster der Vorhersagegüte: Die Modelle liefern im Kurzfristbereich (ca. 24-48 h) die höchsten Gütewerte, während ab Vorlaufzeiten ≥ 72 h ein deutlicher Leistungsabfall einsetzt. Dieses Verhalten ist in den deterministischen Massen ebenso sichtbar wie in den probabilistischen Metriken. Zwischen den Modellgenerationen bestehen jedoch relevante Unterschiede: Die jüngsten Cosmo-Modelle, insbesondere Cosmo-2E, zeigen bis zur Schwelle von $40 \text{ m}^3/\text{s}$ über den gesamten Fünf-Tage-Zeitraum hinweg positive BSS-Werte und damit einen durchgängigen probabilistischen Mehrwert gegenüber der Klimatologie. Die Icon-Modelle weisen zwar kurzfristig eine konkurrenzfähige deterministische Güte und anfänglich positive CRPSS-Werte auf, verlieren diesen probabilistischen Vorteil jedoch rasch: Je nach Schwelle wird die BSS-Nulllinie bereits zwischen etwa 18 und 78 h erreicht oder unterschritten, bei $40 \text{ m}^3/\text{s}$ sogar besonders früh. Die Präzision der Prognosen wird in erster Linie durch die Vorlaufzeit und die Schwellenhöhe geprägt, zusätzlich aber auch durch Kalibrierung und Ensemble-Dispersion (Rang-Histogramme, Attributdiagramme) sowie, für den CRPSS, durch die modell- und zeitraumspezifische Wahl der Klimatologie als Referenz. Die Vergleichbarkeit zwischen den Cosmo- und Icon-Modellen ist insofern eingeschränkt, als die Icon-Zeitreihe in dieser Studie nur rund 1,5 Jahre umfasst, während die Cosmo-Modellen über deutlich längere Zeiträume evaluiert wurde.

Mit Blick auf Fehlerquellen und Unsicherheiten bestätigt die Analyse mehrere robuste Befunde: (i) eine leichte systematische Unterschätzung der Abflüsse, die mit der Vorlaufzeit zunimmt, (ii) eine abnehmende Diskriminationsfähigkeit mit längerer Vorlaufzeit und höheren Schwellen (flacher werdende ROC, sinkende AUC), (iii) Kalibrierungsdefizite und Underdispersion insbesondere bei den Icon-Modellen, sichtbar in U-förmigen Rang-Histogrammen sowie in Attributdiagrammen mit Overforecasting bei kurzen und Underconfidence bei langen Vorlaufzeiten und (iv) ein BSS, der bei Cosmo-2E und Cosmo-1E bis $40 \text{ m}^3/\text{s}$ stabil positiv bleibt, während die Icon-Modelle bei derselben Schwelle schon ab ca. 18 h keinen Mehrwert mehr gegenüber der Klimatologie bietet. Insgesamt sprechen die Kalibrierungsdiagnostik und die Schwellenmetriken dafür, dass Cosmo-2E im operationell relevanten Bereich bis $40 \text{ m}^3/\text{s}$ über die gesamte Laufzeit zuverlässiger nutzbare Wahrscheinlichkeiten lieferte, während Icon-2 seinen Mehrwert stärker auf den sehr kurzfristigen Bereich konzentriert.

Ökonomisch bedeutsam ist, dass positive BSS/CRPSS-Werte verlässlichere, entscheidungsrelevante Information gegenüber der Klimatologie anzeigen. Hieraus ergab sich unter Cosmo-2E bis 40 m³/s über fünf Tage ein stabiler Nutzen für Speicherbewirtschaftung und präventive Einsatzplanung. Seit der Ablösung durch die Icon-Modelle hat sich das Bild verschoben: Kurzfristig ($\approx$ bis 18 h) bleibt die Informationsqualität positiv, aber mittel- und langfristig sinkt der probabilistische Mehrwert jedoch, parallel zu den beobachteten Kalibrierungsdefiziten und der Unterdispersion, deutlich. Am Limmatwehr in Zürich verschärft dies das Kosten-Risikodilemma: Vorsorgliche Schliessungen bei ausbleibender Schwellenüberschreitung verursachen Stillstand, Verzögerungen und Zusatzkosten, während unterlassene Massnahmen bei tatsächlich eintretender Überschreitung zu Schäden und Sicherheitsrisiken führen. Je geringer die Zuverlässigkeit der Wahrscheinlichkeitsinformation ab Tag 2, desto höher der Interpretations- und Abklärungsaufwand und desto grösser das Risiko ökonomisch ineffizienter Entscheidungen.

Für endnutzerrelevante Schwellen, insbesondere 40 m³/s, bedeutet dies: Die Erkennungsleistung nimmt mit Vorlaufzeit deutlich ab. Operativ sollten Icon-basierte, schwellenorientierte Entscheidungen deshalb stärker auf den Kurzfristbereich fokussieren und ab Tag 2 konsequent durch statistische Nachkalibrierung ergänzt werden.

Aus diesen Befunden ergeben sich vier prioritäre Schritte für den Ausblick: Erstens sollte eine einheitliche, langfristige Klimatologie als gemeinsame Referenz für alle Modelle etabliert werden, um CRPSS-Vergleiche zu objektivieren. Zweitens ist die Zeitreihe der Icon-Modelle zu verlängern, damit interannuelle Stabilität und Extremfall-Leistung belastbar beurteilt werden können. Drittens sind statistische Nachkalibrierungen sowie gezielte Massnahmen gegen Unterdispersion einzusetzen, um BSS/CRPSS jenseits von 24 h zu stabilisieren. Viertens empfiehlt sich für das Limmatwehr eine formale, projektspezifische Cost-Loss-Analyse mit konsistenter Bewertungsperiode und klaren, probabilistisch definierten Einsatzregeln, damit der operative und ökonomische Nutzen der Vorhersagen, insbesondere unter den Icon-Modelle, transparent maximiert werden kann. Die Kernbotschaft lautet: Icon-2 liefert kurzfristig einen Nutzen, mittelfristig ist ohne Nachkalibrierung mit einem deutlichen Mehrwertverlust zu rechnen. Eine längere Datenbasis, einheitliche Referenz, systematische Kalibrierung und entscheidungsorientierte Bewertung sind die Hebel, um den praktischen Nutzen im Betrieb wieder signifikant zu erhöhen.

8. Literaturverzeichnis

- Addor, N., Jaun, S., Fundel, F., Zappa, M., 2011. An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* 15, 2327–2347. <https://doi.org/10.5194/hess-15-2327-2011>
- AWEL, 2017. Hochwasserschutz an Sihl, Zürichsee und Limmat - Gefährdung und Massnahmen im Überblick. Amt für Abfall, Wasser, Energie und Luft (AWEL), Zürich.
- AWEL, 2019. Erneuerung Platzspitzwehr. Amt für Abfall, Wasser, Energie und Luft (AWEL), Zürich.
- Badoux, A., Zappa, M., Oplatka, M., Bösch, M., Jaun, S., Steiner, P., Hegg, C., Rhyner, J., 2010. IFKIS-Hydro Sihl: Beratung und Alarmorganisation während des Baus der Durchmesserlinie beim Hauptbahnhof Zürich. 4 102, 309–320.
- Bowers, A.J., Zhou, X., 2019. Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *J. Educ. Stud. Placed Risk JESPAR* 24, 20–46. <https://doi.org/10.1080/10824669.2018.1523734>
- Braun, L.N., 1984. Simulation of snowmelt-runoff in lowland and lower alpine regions of Switzerland. *Proceedings of the Buda pest Symposium on Modelling Snowmelt-Induced Processes*, July 1986. <https://doi.org/10.3929/ETHZ-A-000334295>
- Braun, L.N., Aellen, M., 1990. Modelling discharge of glacierized basins assisted by direct measurements of glacier mass balance, in: H. Lang & A. Musy (Eds.), *Hydrology in Mountainous Regions. I—Hydrological Measurements*, IAHS Publ. IAHS, pp. 99–106.
- Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Softw.* 25, 854–872. <https://doi.org/10.1016/j.envsoft.2010.01.009>
- Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. *J. Hydrol.* 375, 613–626. <https://doi.org/10.1016/j.jhydrol.2009.06.005>
- Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. <https://doi.org/10.1175/BAMS-D-12-00081.1>
- Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., Ramos, M.-H., 2007. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environ. Hazards, Environmental Hazards and Risk Communication* 7, 115–127. <https://doi.org/10.1016/j.envhaz.2007.05.001>
- Etzelwerk AG, 2017. Stauanlage Sihlsee – Wehrrglement (Version 0.3, April 2017). Altendorf/Zollikofen.
- Horton, P., Schaeffli, B., Kauzlaric, M., 2022. Why do we have so many different hydrological models? A review based on the case of Switzerland. *WIREs Water* 9, e1574. <https://doi.org/10.1002/wat2.1574>
- Hottelet, C., Braun, L.N., Leibundgut, C., Rieg, A., 1992. Simulation of Snowpack and Discharge in an Alpine Karst Basin, in: *Proceedings of the Kathmandu Symposium on Snow and Glacier Hydrology*, IAHS Publ. IAHS.
- Kleinn, J., Aller, D., Zappa, M., Bresch, D., Marti, C., Oplatka, M., 2019. Kontinuierliche Wirkungsabschätzung von verschiedenen Hochwasserschutzmassnahmen über das gesamte Abfluss-Spektrum am Beispiel der Sihl. *Hydrologie & Wasserbewirtschaftung* 63, 158–167. https://doi.org/10.5675/HYWA_2019.3_3
- Liechti, K., Oplatka, M., Eisenhut, N., Zappa, M., 2016. Early Flood Warning for the City of Zurich: Evaluation of real-time Operations since 2010. *Intrapraevent* 2016 944–951.
- Liechti, K., Zappa, M., 2012. Quantitative Auswertung des Sihl Vorhersagesystems II: 02/2010-09/2012, Unveröffentlichter Bericht im Auftrag des AWEL Zürich. Eidgenössische Forschungsanstalt WSL, Brimensdorf.

- Liechti, K., Zappa, M., 2014. Quantitative Auswertung des Sihl Vorhersagesystems III: Februar 2010–Oktober 2014, Unveröffentlichter Bericht im Auftrag des AWEL Zürich. Eidgenössische Forschungsanstalt WSL, Brimensdorf.
- Liechti, K., Zappa, M., 2017. Sihl 2016: Auswertungen Oktober 2015–September 2016, Unveröffentlichter Bericht im Auftrag des AWEL Zürich. Eidgenössische Forschungsanstalt WSL, Brimensdorf.
- Liechti, K., Zappa, M., 2020. Sihl 2019: Auswertungen Oktober 2015 – Oktober 2019, Unveröffentlichter Bericht im Auftrag des AWEL Zürich. Eidgenössische Forschungsanstalt WSL, Brimensdorf.
- Lugiez, F., Kasser, P., Jensen, H., Guillot, P., 1969. La prévision des débits du Rhin. *Houille Blanche* 55, 733–746. <https://doi.org/10.1051/lhb/1969058>
- Martinez, J., 1970. Study of snowmelt runoff process in two representative watersheds with different elevation range, in: Symposium on the Results of Research in Representative and Experimental Basins, IAHS Publ. IAHS, Wellington, pp. 29–39.
- MeteoSchweiz, 2025a. ICON-Prognosesysteme [WWW Document]. Bundesamt Für Meteorol. Klimatol. MeteoSchweiz. URL <https://www.meteoschweiz.admin.ch/wetter/warn-und-prognosesysteme/icon-prognosesysteme.html> (accessed 10.8.25).
- MeteoSchweiz, 2025b. Das neue ICON Wettermodell. Bundesamt für Meteorologie und Klimatologie.
- Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.* 10, 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Rossa, A., Liechti, K., Zappa, M., Bruen, M., Germann, U., Haase, G., Keil, C., Krahe, P., 2011. The COST 731 Action: A review on uncertainty propagation in advanced hydro-meteorological forecast systems. *Atmospheric Res., Uncertainty Propagation in Advanced Hydro-Meteorological Forecast Systems* 100, 150–167. <https://doi.org/10.1016/j.atmosres.2010.11.016>
- Schwanbeck, J., Viviroli, D., Weingartner, R., Röser, I., Trösch, J., 2007. Prozessbasierte Abschätzung von Hochwassern im Einzugsgebiet der Sihl. Schlussbericht zur Studie im Auftrag des Amtes für Abfall, Wasser, Energie und Luft (AWEL) des Kantons Zürich.
- Thielen, J., Schaake, J., Hartman, R., Buizza, R., 2008. Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. *Atmospheric Sci. Lett.* 9, 29–35. <https://doi.org/10.1002/asl.168>
- Troin, M., Arsenault, R., Wood, A.W., Brissette, F., Martel, J.-L., 2021. Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years. *Water Resour. Res.* 57, e2020WR028392. <https://doi.org/10.1029/2020WR028392>
- Weigel, A.P., Liniger, M.A., Appenzeller, C., 2007. The Discrete Brier and Ranked Probability Skill Scores. <https://doi.org/10.1175/MWR3280.1>
- Wilks, D.S., 2006. Statistical methods in the atmospheric sciences, 2nd ed. ed, International geophysics series. Academic Press, Amsterdam ; Boston.
- World Meteorological Organization (WMO), 1975. Intercomparison of conceptual models used in operational hydrological forecasting (No. 7), Operational hydrology report (OHR). World Meteorological Organization (WMO), Geneva.
- Zappa, M., Jaun, S., Badoux, A., Schwanbeck, J., Addor, N., Liechti, K., Roeser, I., Walser, A., Viviroli, D., Vogt, S., Gerber, M., Trösch, J., Weingartner, R., Oplatka, M., Bezzola, G.R., Rhyner, J., 2010. IFKIS-Hydro Sihl: Ein operationelles Hochwasservorhersagesystem für die Stadt Zürich und das Sihltal. *Wasser Energ. Luft*, 3 102, 238–248.
- Zappa, M., Jaun, S., Germann, U., Walser, A., Fundel, F., 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Res., Uncertainty Propagation in Advanced Hydro-Meteorological Forecast Systems* 100, 246–262. <https://doi.org/10.1016/j.atmosres.2010.12.005>
- Zappa, M., Rotach, M.W., Arpagaus, M., Dörninger, M., Hegg, C., Montani, A., Ranzi, R., Ament, F., Germann, U., Grossi, G., Jaun, S., Rossa, A., Vogt, S., Walser, A., Wehrhan, J., Wunram, C., 2008. MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems. *Atmospheric Sci. Lett.* 9, 80–87. <https://doi.org/10.1002/asl.183>

Zappa, M., Vogt, S., 2007. Hochwasser-Vorhersagesysteme der neusten Generation im Praxis-Test, in: C. Hegg & J. Rhyner (Eds.), Forum für Wissen: Vol. 2007. Warnung bei aussergewöhnlichen Naturereignissen, Forum für Wissen. Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft (WSL), Brimensdorf, pp. 25–31.

9. Anhang

9.1 Formeln und Definitionen der verwendeten Gütemassen

Folgende Definitionen basieren auf Wilks (2006), sofern nicht anders vermerkt. y beschreibt die Prognosewerte und o beschreibt die Beobachtungswerte.

9.1.1 Deterministische, kontinuierliche Gütemasse

9.1.1.1 **Mean Error (ME)** - Mittlerer Fehler

«Der mittlere Fehler (ME) ist [...] der Mittelwert der Differenzen zwischen Vorhersagen und Beobachtungen. Er ist ein Mass für die Verzerrung.» (Wilks, 2006)

Der ME gibt an, ob ein Modell im Durchschnitt über- oder unterschätzt. Wenn der ME grösser als 0 ist, überschätzt das Modell systematisch, wenn ME kleiner als 0 unterschätzt das Modell systematisch und wenn ME gleich 0 ist, gibt es im Mittel keine Verzerrung.

Formel 1

$$ME = \frac{1}{n} \sum_{k=1}^n (y_k - o_k) = \bar{y} - \bar{o}$$

9.1.1.2 **Mean Absolute Error (MAE)** - Mittlere absolute Fehler

Der mittlere absolute Fehler beschreibt den Durchschnitt der absoluten Differenzen zwischen vorhergesagten und beobachteten Werten. Ein kleiner MAE bedeutet, dass das Modell im Schnitt nahe den echten Werten liegt. Es ist robust gegenüber Ausreisser, nimmt jedoch keine Rücksicht auf die Richtung des Fehlers.

Formel 2

$$MAE = \frac{1}{n} \sum_{k=1}^n |y_k - o_k|$$

9.1.1.3 **Root Mean Square Error (RMSE)** - Wurzel des mittleren quadratischen Fehlers

Die Wurzel des mittleren quadratischen Fehlers quantifiziert die Abweichung zwischen Modellvorhersagen und tatsächlichen Beobachtungen. Ein kleiner RMSE-Wert bedeutet, dass die Vorhersagen im Schnitt nahe an den beobachteten Werten liegen. RMSE ist im Gegensatz zu MAE anfällig für Ausreisser, da durch die Quadrierung grosse Fehler stärker gewichtet werden.

Formel 3

$$RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^n (y_k - o_k)^2}$$

9.1.1.4 **Pearson's Correlation Coefficient r** – Pearson-Korrelationskoeffizient

«Der Pearson-Korrelationskoeffizient r misst die Stärke der linearen Beziehung zwischen den Vorhersagen und den Beobachtungen.» (Wilks, 2006)

Der Korrelationskoeffizient r kann einen Wert zwischen -1 und 1 annehmen, wobei r nahe 1 auf eine starke positive, r nahe -1 auf eine stark negative lineare Beziehung hinweist. Wenn r nahe 0 ist, deutet das auf keine lineare Beziehung zwischen Vorhersage und Beobachtung.

Formel 4

$$r = \frac{\sum_{k=1}^n (y_k - \bar{y})(o_t - \bar{o})}{\sqrt{\sum_{k=1}^n (y_k - \bar{y})^2} \sqrt{\sum_{k=1}^n (o_t - \bar{o})^2}}$$

9.1.1.5 **Coefficient of determination R^2** - Bestimmtheitsmass

Das Bestimmtheitsmass R^2 beschreibt, wie gut eine Regression die Gesamtvarianz zwischen zwei Variablen erfasst. Es wird definiert als das Verhältnis der durch das Regressionsmodell erklärten Varianz (SSR) zur gesamten beobachteten Varianz (SST). R^2 gibt somit den Anteil der Gesamtstreuung der abhängigen Variable an, der durch das Modell erklärt wird. Ein Wert von $R^2 = 1$ kennzeichnet eine perfekte Anpassung, während $R^2 = 0$ bedeutet, dass das Modell keine erklärende Kraft besitzt. Bei einer einfachen linearen Regression entspricht die Wurzel von R^2 dem Pearson-Korrelationskoeffizienten zwischen den beiden Variablen.

Formel 5

$$R^2 = 1 - \frac{SSE}{SST}$$

9.1.1.6 **Nash–Sutcliffe-Efficiency (NSE)** - Nash–Sutcliffe-Effizienz

Die Nash–Sutcliffe-Effizienz vergleicht den mittleren quadratischen Fehler der Vorhersage mit der Varianz der Beobachtungen. Die NSE-Werte reichen von $-\infty$ zu 1. Ein Wert von 1 steht für eine perfekte Übereinstimmung der Vorhersagen und Beobachtungen, während ein Wert von 0 bedeutet, dass die Modellvorhersagen gleichgut wie der Mittelwert der Beobachtungen sind. Negative NSE-Werte zeigen an, dass die beobachteten Mittelwerte eine bessere Schätzung darstellen als das Modell. (Nash and Sutcliffe, 1970)

Formel 6

$$NSE = 1 - \frac{\sum_t (y_t - o_t)^2}{\sum_t (o_t - \bar{o})^2}$$

9.1.2 Kontingenztafelbasierte Metriken

Zur Bewertung der Güte binärer Vorhersagen (Ereignis / kein Ereignis) wird üblicherweise eine 2x2-Kontingenztafel herangezogen. Sie zeigt, wie häufig jede Kombination aus Vorhersage und tatsächlichem Ergebnis auftritt und bildet damit die Grundlage für viele wichtige Verifikationskennzahlen. In der Tabelle werden vier Fälle unterschieden: Ein Ereignis wird korrekt

vorhergesagt (Hit, «H»), ein Ereignis wird vorhergesagt, obwohl es nicht eintritt (False Alarm, «F»), ein Ereignis tritt ein, obwohl es nicht vorhergesagt wurde (Miss, «M»), und schliesslich wird korrekt erkannt, dass kein Ereignis stattfindet (Correct Negative, «C»). Die Kontingenztabelle zeigt somit die absolute Häufigkeitsverteilung aller möglichen Kombinationen aus Prognosen und Beobachtungen. Teilt man jede Zelle durch die Gesamtzahl der Fälle (n), so ergeben sich relative Häufigkeiten, die der gemeinsamen Verteilung von Vorhersage und Beobachtung entsprechen. Auf Basis dieser Werte lassen sich verschiedene Masszahlen ableiten, darunter die Probability of Detection (POD), die False Alarm Ratio (FAR) sowie die Probability of False Detection (POFD), die zusammen eine detaillierte Beurteilung der Vorhersagequalität ermöglichen.

Table 1: 2 x 2 Kontingenztabelle.

		Beobachtung		Summe Vorhersagen
		Ereignis eingetreten	Ereignis nicht eingetreten	
Vorhersage	Ereignis	H (Hit)	F (False Alarm)	H + F
	kein Ereignis	M (Miss)	C (Correct Negative)	M + C
Summe Beobachtungen		H + M	F + C	n

9.1.2.1 **Probability of Detection (POD) - Erkennungswahrscheinlichkeit**

Die Erkennungswahrscheinlichkeit bezeichnet die Anzahl Fälle, in denen das prognostizierte Ereignis tatsächlich eingetreten ist.

Formel 7

$$\text{POD} = \frac{H}{H + M}$$

9.1.2.2 **False Alarm Ratio (FAR)**

Der FAR gibt an, wie viele der als positiv vorhergesagten Fälle sich als falsch herausstellen, also wie gross der Anteil der prognostizierten Ereignisse ist, die tatsächlich nicht eintreten. Durch die negative Ausrichtung sind kleinere FAR-Werte vorzuziehen. Der ideale FAR liegt bei 0, während der schlechteste Wert 1 beträgt.

Formel 8

$$\text{FAR} = \frac{F}{H + F}$$

9.1.2.3 **Probability of False Detection (POFD)**

Die POFD beschreibt die Wahrscheinlichkeit, dass ein Ereignis fälschlicherweise als eingetreten erkannt wird, obwohl es tatsächlich nicht eingetreten ist.

Formel 9

$$\text{POFD} = \frac{F}{F + C}$$

9.1.2.4 **Critical Success Index (CSI)**

Der CSI misst den Anteil der korrekt vorhergesagten Ereignisse (Hits) an allen Fällen, in denen das Ereignis entweder vorhergesagt oder tatsächlich beobachtet wurde. Einen CSI von 1 bedeutet eine perfekte Vorhersage, alle Ereignisse wurden korrekt erkannt ohne Fehlalarme oder verpasst Fälle. Bei einem CSI von 0 gibt es keine Übereinstimmung zwischen Vorhersagen und Beobachtungen.

Formel 10

$$\text{CSI} = \frac{H}{H + M + F}$$

9.1.2.5 **Heidke Skill Score (HSS)**

Der Heidke Skill Score (HSS) bewertet die Gesamtgüte einer Vorhersage im Vergleich zu einer zufälligen (reinen Glücks-)Vorhersage. Der HSS zeigt, wie viel besser ein Vorhersagesystem im Mittel arbeitet, als man es durch Zufall erwarten würde. Ein Wert von 0 bedeutet, dass die Vorhersage keine Verbesserung gegenüber einer zufälligen Prognose darstellt. Negative Werte des HSS zeigen, dass das Vorhersagesystem schlechter abschneidet als eine zufällige Vorhersage.

Formel 11

$$\text{HSS} = \frac{2(HC - MF)}{(H + M)(M + C) + (H + F)(F + C)}$$

9.1.3 Probabilistische Metriken für binäre Ereignisse

9.1.3.1 **Brier Score (BS)**

Der Brier Score beschreibt den mittleren quadratischen Fehler probabilistischer Vorhersagen. Es bewertet die Genauigkeit der probabilistischen Vorhersagen, also solche, die Wahrscheinlichkeiten für das Eintreten eines Ereignisses angeben (zum Beispiel «die Wahrscheinlichkeit, dass die Sihl einen Abfluss von 40 m³/s erreicht, beträgt 70%»). Dabei wird die Beobachtung mit $o=1$ codiert, wenn das Ereignis tatsächlich eintritt, und mit $o=0$, wenn es ausbleibt. Der Score ergibt sich als Mittelwert der quadrierten Abweichungen zwischen den vorhergesagten Wahrscheinlichkeiten und den entsprechenden binären Beobachtungen. Ein BS von 0 bedeutet eine perfekte Prognose und 1 eine komplett falsche Prognose.

Formel 12

$$\text{BS} = \frac{1}{n} \sum_{t=1}^n (y_t - o_t)^2$$

9.1.3.2 **Brier Skill Score (BSS)**

Der Brier Skill Score ist eine standardisierte Form des Brier Scores und bewertet, wie viel besser oder schlechter eine Wahrscheinlichkeitsvorhersage im Vergleich zu einer Referenzvorhersage (meist Klimatologie) ist. So unterscheidet sich der BSS vom BS durch das in Relation zu einer Referenz setzen. Ein BSS von 1 drückt eine perfekte Vorhersage aus, ein BSS von 0 bedeutet, dass die Vorhersage keine Verbesserung gegenüber einer Referenz aufweist und ein BSS kleiner als 0 zeigt eine schlechtere Vorhersage als wenn man die Referenzvorhersage nutzen würde.

Formel 13

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}$$

9.1.3.3 **Receiver Operating Characteristic (ROC) Curve – Grenzwertoptimierungskurve**

Das ROC-Diagramm dient zur grafischen Beurteilung der Fähigkeit eines Vorhersagesystems, zwischen dem Eintreten und dem Nichteintreten eines Ereignisses zu unterscheiden. Auf der y-Achse wird POD (Kapitel 9.1.2.1) dargestellt, während auf der x-Achse POFD (Kapitel 9.1.2.3) aufgetragen wird. Jeder Punkt der Kurve entspricht einer bestimmten Entscheidungsschwelle, ab der ein Ereignis als eingetreten gilt. Eine rein zufällige Vorhersage würde als Diagonale im Diagramm erscheinen (POD = POFD), während eine gut trennende Vorhersage eine Kurve zeigt, die möglichst weit nach oben links verläuft. Je stärker die Kurve von der Diagonalen abweicht, desto höher ist die Trennschärfe des Modells. Als zusammenfassendes Mass wird häufig die Fläche unter der Kurve (AUC – Kapitel 9.1.3.4) verwendet, die Werte zwischen 0.5 (zufällige Trennung) und 1 (perfekte Trennung) annehmen kann.

9.1.3.4 **Area Under the Curve (AUC) – Fläche unter der Kure**

Die AUC beschreibt die Fläche unter der ROC-Kurve (Kapitel 9.1.3.3) und dient als zusammenfassendes Mass für die Fähigkeit eines Vorhersagesystems, zwischen dem Eintreten und dem Nichteintreten eines Ereignisses zu unterscheiden. Eine AUC von 1 steht für eine perfekte Trennung, während eine AUC von 0.5 einer zufälligen Vorhersage entspricht. Werte unter 0.5 deuten darauf hin, dass das Modell schlechter abschneidet als Zufall, also Ereignisse und Nicht-Ereignisse systematisch vertauscht. Damit misst sie die Diskriminationsfähigkeit des Modells, unabhängig von der gewählten Entscheidungsschwelle.

Formel 14

$$\text{AUC} = \int_0^1 f(x) dx$$

(Bowers and Zhou, 2019)

9.1.3.5 **Reliability-Diagramm – Reliabilitätsdiagramm**

Das Reliabilitätsdiagramm bewertet die Kalibrierung einer probabilistischen Vorhersage. Es zeigt, inwieweit die vorhergesagten Wahrscheinlichkeiten mit den tatsächlich beobachteten Eintrittshäufigkeiten übereinstimmen. Auf der x-Achse werden die vorhergesagten Wahrscheinlichkeiten dargestellt, während die y-Achse die beobachteten relativen Häufigkeiten angibt. Für die Berechnung werden die Vorhersagen in Wahrscheinlichkeitsklassen unterteilt, und für jede Klasse wird bestimmt, wie oft das Ereignis tatsächlich eingetreten ist. Ein perfekt kalibriertes Modell liegt auf der Diagonalen ($y = x$), da in diesem Fall eine Prognose von 70 % Eintrittswahrscheinlichkeit

bedeutet, dass das Ereignis auch in etwa zu 70 % auftritt. Punkte unterhalb der Diagonalen deuten auf eine Überschätzung der Eintrittswahrscheinlichkeit hin, während Punkte oberhalb eine Unterschätzung anzeigen.

9.1.3.6 **Attribute Diagramm – Attributdiagramm**

Ein Attributdiagramm ist eine erweiterte Form des Reliabilitätsdiagramms, welches zusätzlich Bezugslinien aus der algebraischen Zerlegung des Brier-Scores (Kapitel 9.1.3.1) und des Brier Skill Score (Kapitel 9.1.3.2) enthält. Es ermöglicht eine umfassende Beurteilung der Zuverlässigkeit (Reliability), des Unterscheidungsvermögens beziehungsweise der Auflösung (Resolution) und der Vorhersagegüte (Skill) probabilistischer Prognosen. Die Zerlegung des Brier Scores erfolgt nach

$$BS = Reliability - Resolution + Uncertainty$$

wobei der Brier Skill Score definiert ist als

$$BSS = \frac{Resolution - Reliability}{Uncertainty}$$

Ein positiver Skill liegt vor, wenn das Modell eine höhere Auflösung als Kalibrierungsfehler aufweist, also wenn Resolution grösser als Reliability ist. Im Attributdiagramm stellen die Perfect-Reliability-Linie, die No-Resolution-Linie (auf Höhe der Klimatologie) und die No-Skill-Linie grafische Referenzen dar, anhand derer beurteilt werden kann, ob eine Vorhersage nützliche Information über die Klimatologie hinaus liefert. Punkte oberhalb der No-Skill-Linie tragen positiv zur Vorhersagegüte bei, während Punkte darunter keinen oder negativen Skill anzeigen.

9.1.4 Probabilistische Metriken für kontinuierliche Grössen

9.1.4.1 **Continuous Ranked Probability Score (CRPS)**

Der Continuous Ranked Probability Score (CRPS) misst die Übereinstimmung zwischen einer Vorhersage und einer Beobachtung in einer kontinuierlichen Skala. Der CRPS berücksichtigt nicht nur den Unterschied zwischen dem Vorhersagewert und der Beobachtung, sondern auch die Unsicherheit und die Form der Vorhersageverteilung. Die negative Ausrichtung des CRPS's bedeutet, dass ein Modell die Beobachtung besser vorhersagen konnte, wenn der CRPS-Wert kleiner ist.

Formel 15

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 dy$$

$$F_o(y) = \begin{cases} 0, & y < \text{beobachteter Wert} \\ 1, & y \geq \text{beobachteter Wert} \end{cases}$$

9.1.4.2 **Continuous Ranked Probability Skill Score (CRPSS)**

Der CRPSS ist eine skalierte Version des CRPS, die den Vorhersagefehler im Verhältnis zu einer Referenzvorhersage bewertet. Er ermöglicht es, die Fähigkeit eines Modells zu bewerten, indem es mit einer Baseline-Vorhersage (zum Beispiel Klimatologie) verglichen wird. Der CRPSS-Wert kann Werte zwischen $-\infty$ und 1 annehmen. Das Modell ist perfekt, wenn der CRPSS-Wert 1 ist. Bei einem CRPSS-Wert von 0 ist das Modell genauso gut wie die Referenzvorhersage und bei negativem CRPSS-Wert schneidet das Modell schlechter als die Referenz ab.

Formel 16

$$\text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{ref}}}$$

(Weigel et al., 2007)

9.1.5 **Ensembleverteilung**

9.1.5.1 **Rank Histogramm – Rang-Histogramm**

Ein Rang-Histogramm ist ein Diagramm, um die Qualität probabilistischer Vorhersagen zu bewerten. Es zeigt, wie gut ein Modell in der Lage ist, die wahrscheinlichen Werte eines Ereignisses (zum Beispiel das Überschreiten von 40 m³/s beim Limmatwehr) korrekt einzuordnen, indem es Vorhersagen und tatsächliche Beobachtungen vergleicht. Dabei wird für jede Vorhersage der «Rang» in Bezug auf die tatsächliche Beobachtung ermittelt und die Häufigkeit dieser Rangwerte in einem Histogramm dargestellt. Ein gut kalibriertes Modell führt zu einem gleichmässigen Rang-Histogramm, während Verschiebungen nach rechts oder links darauf hinweisen, dass das Modell zu optimistisch oder zu pessimistisch in seinen Vorhersagen ist.

9.2 **Reproduzierbarkeit**

Die Auswertungen wurden in R (Version 4.5.0) durchgeführt, unter Verwendung der Packages «dplyr», «tidyr», «purrr», «stringr», «lubridate», «ggplot2», «scoringRules», «readr» und «tibble».

Die Verifikation basiert auf die modellierten und beobachteten Zeitreihen, die zunächst station-, modell- und periodenweise in .RData-Dateien zusammengeführt wurden. Jede Datei enthält Ensembleprognosen (cE_rf), Beobachtungen (cE_obs) sowie die zugehörigen Gültigkeitszeiten (cE_validtime). Zur Verarbeitung wurden alle Dateien automatisiert eingelesen und in einer Schleife pro Modell, Station und Jahr analysiert. Zeitangaben wurden dabei in UTC-normalisierte POSIX-Zeitstempel umgewandelt.

9.3 **Erklärung der Nutzung von Künstlicher Intelligenz**

Für die Erstellung dieser Masterarbeit wurde die KI-Plattform ChatGPT-5 von OpenAI eingesetzt. Die Nutzung beschränkte sich auf:

- Code-Generierung, insbesondere zur Unterstützung bei Datenanalysen und Berechnungen.
- Paraphrasierung und sprachliche Überarbeitung eigener Formulierungen.
- Grammatikalische Korrektur der Texte, um Lesbarkeit und sprachliche Präzision zu verbessern.

9.4 Zusatzabbildungen

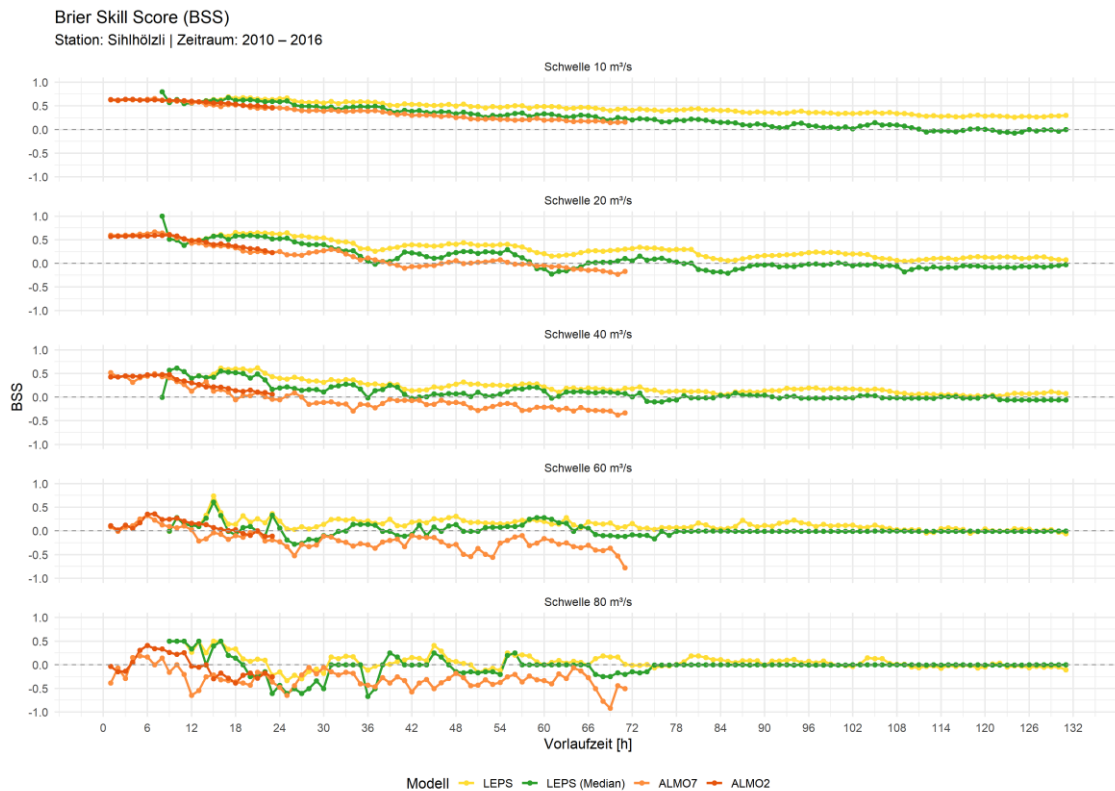


Abbildung 33: Brier Skill Score (BSS) der meteorologischen Modelle Cosmo-Leps, Cosmo-2 und Cosmo-7 für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2010-2016.

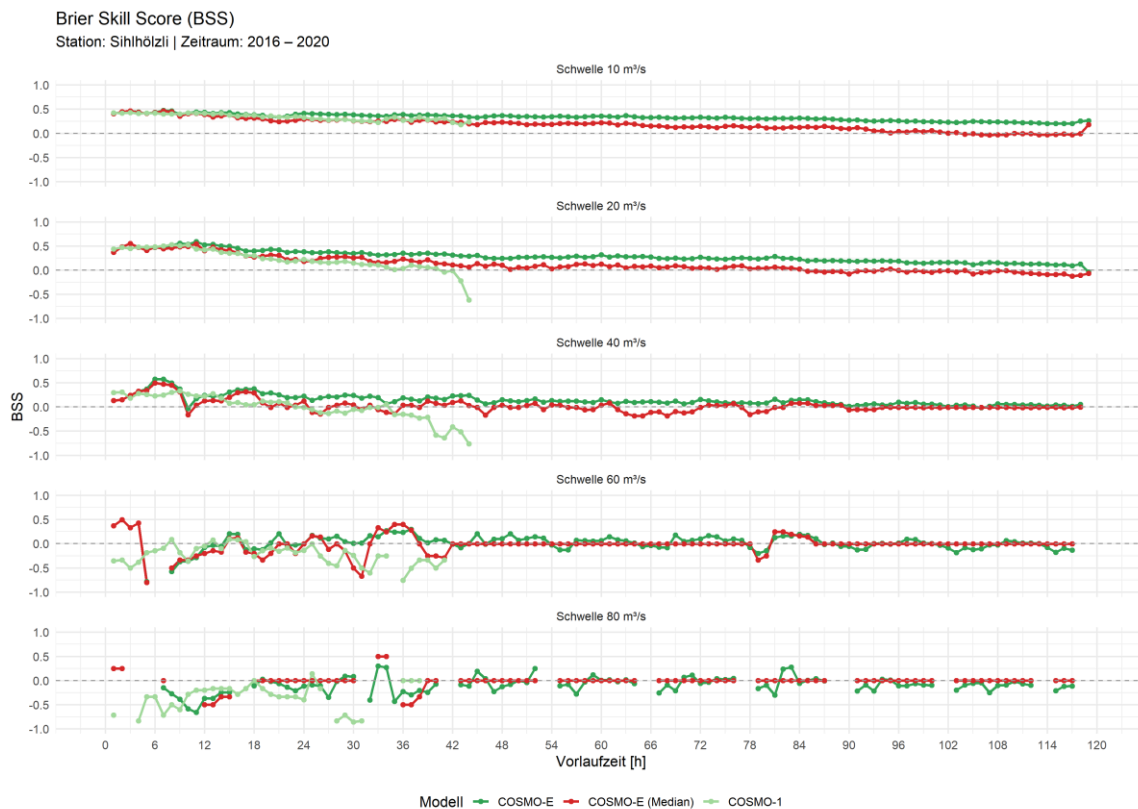


Abbildung 34: Brier Skill Score (BSS) der meteorologischen Modelle Cosmo-1 und Cosmo-E für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2016-2020.

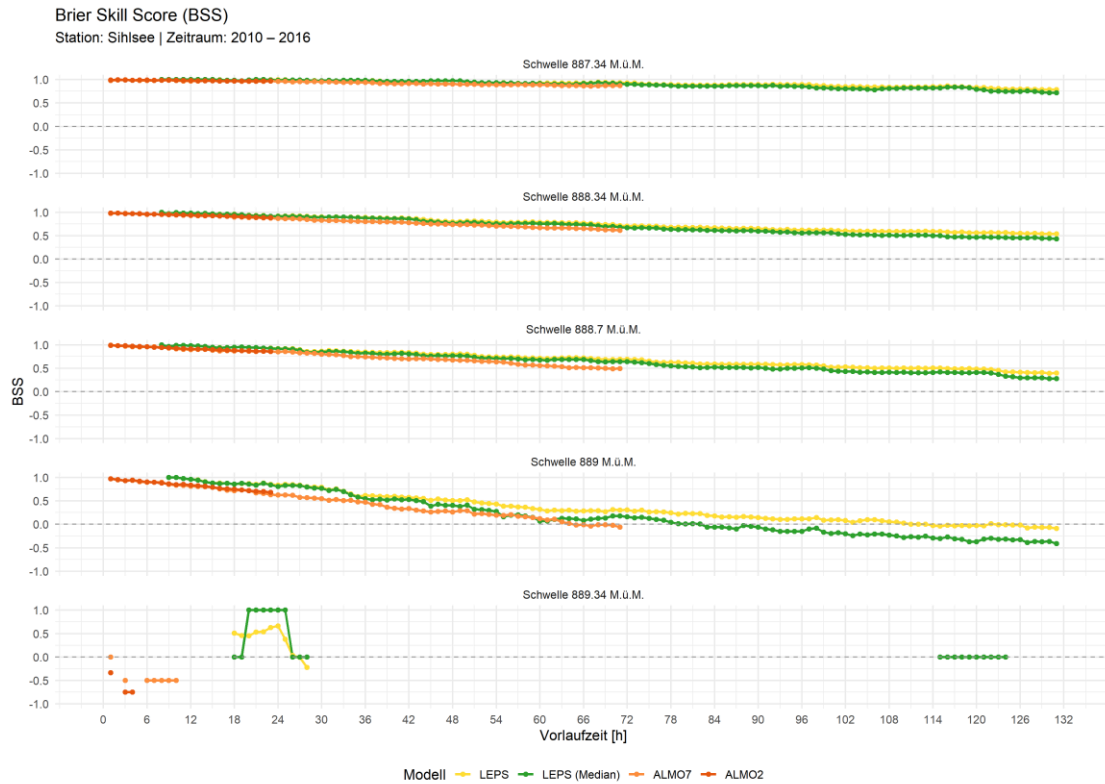


Abbildung 35: Brier Skill Score (BSS) der meteorologischen Modelle Cosmo-Leps, Cosmo-2 und Cosmo-7 für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2010-2016.

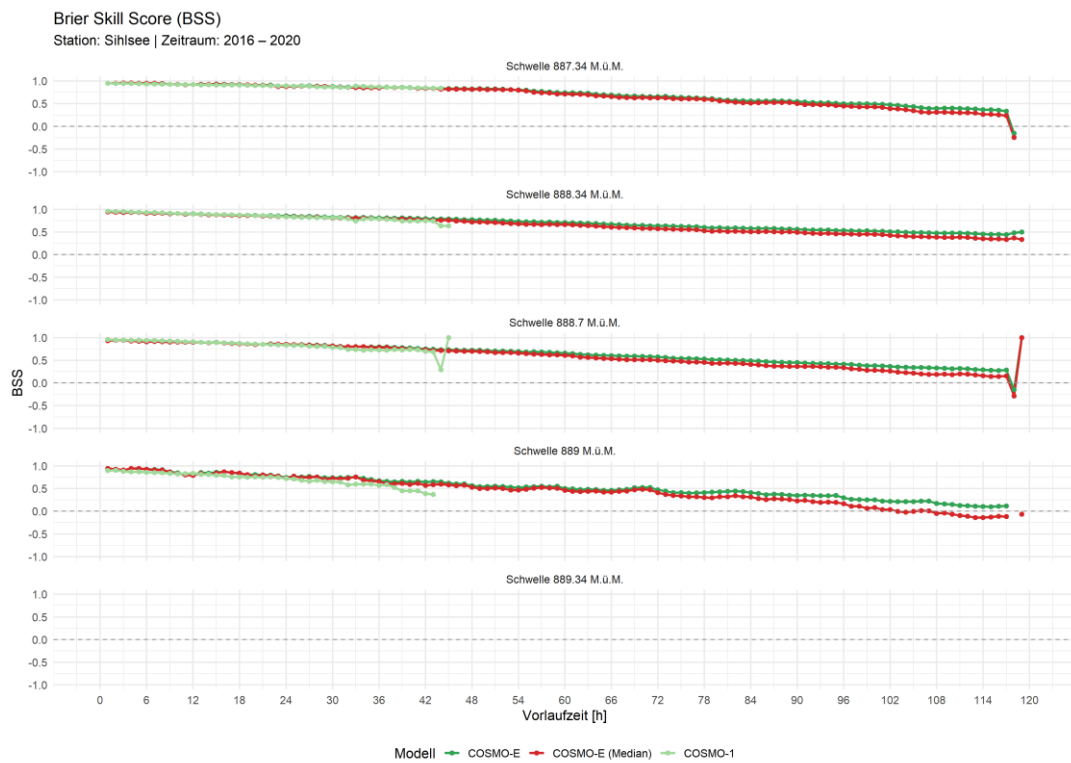


Abbildung 36: Brier Skill Score (BSS) der meteorologischen Modelle Cosmo-1 und Cosmo-E für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2016-2020.

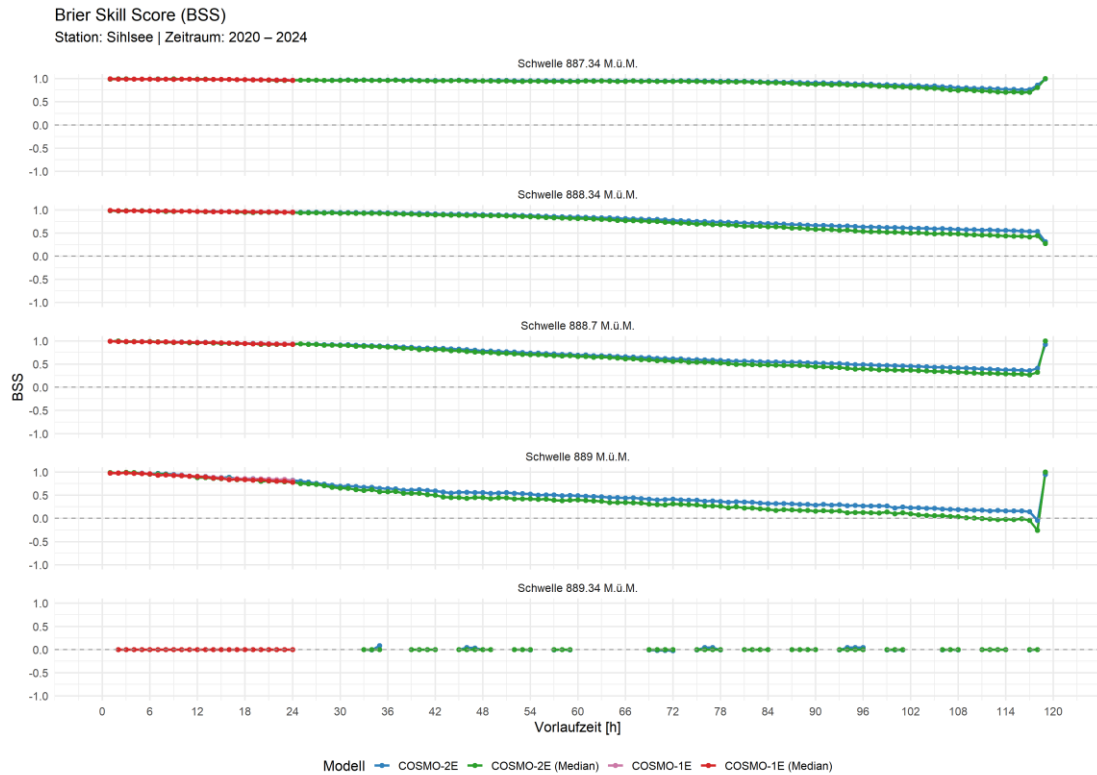


Abbildung 37: Brier Skill Score (BSS) der meteorologischen Modelle Cosmo-1E und Cosmo-2E für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2020-2024.

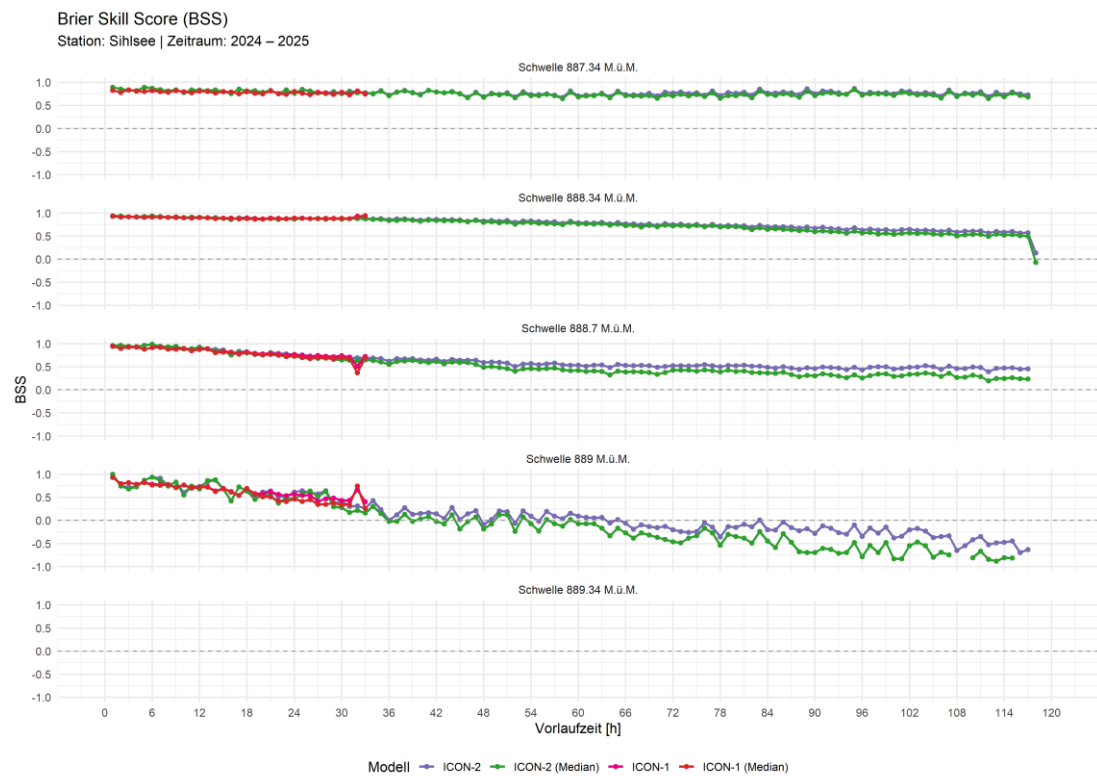


Abbildung 38: Brier Skill Score (BSS) der meteorologischen Modelle Icon-1 und Icon-2 für verschiedene Abflussschwellen und Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2024-2025.

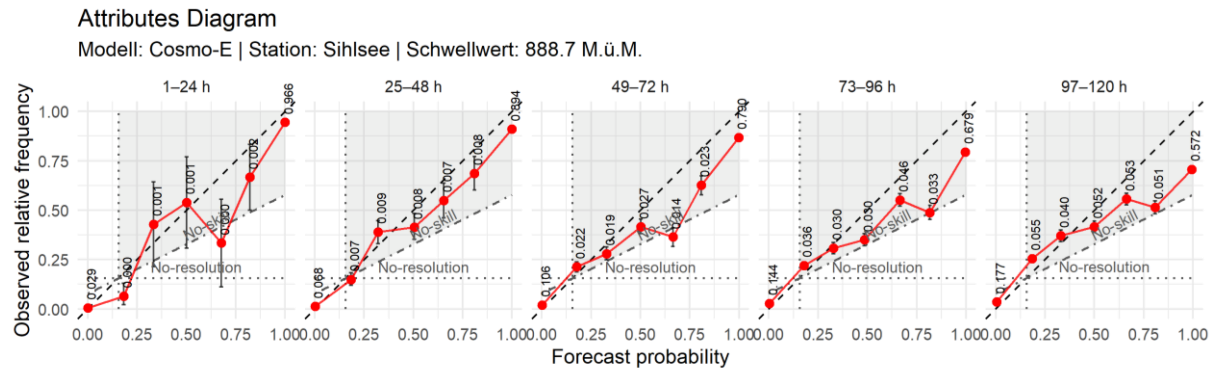


Abbildung 39: Attribut Diagramm des meteorologischen Modells Cosmo-E für den Schwellwert 888.7 m. ü. M. und verschiedene Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2016-2020.

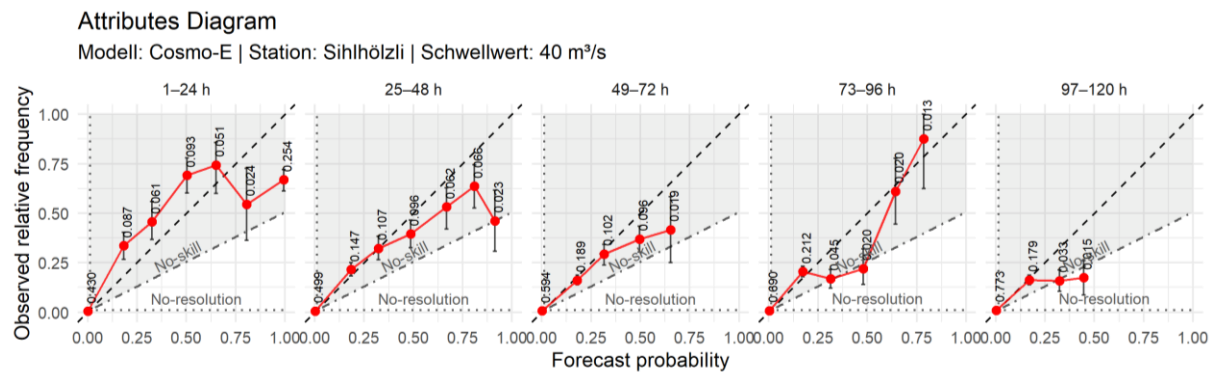


Abbildung 40: Attribut Diagramm des meteorologischen Modells Cosmo-E für die Abflussschwelle von 40 m³/s und verschiedene Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2016-2020.

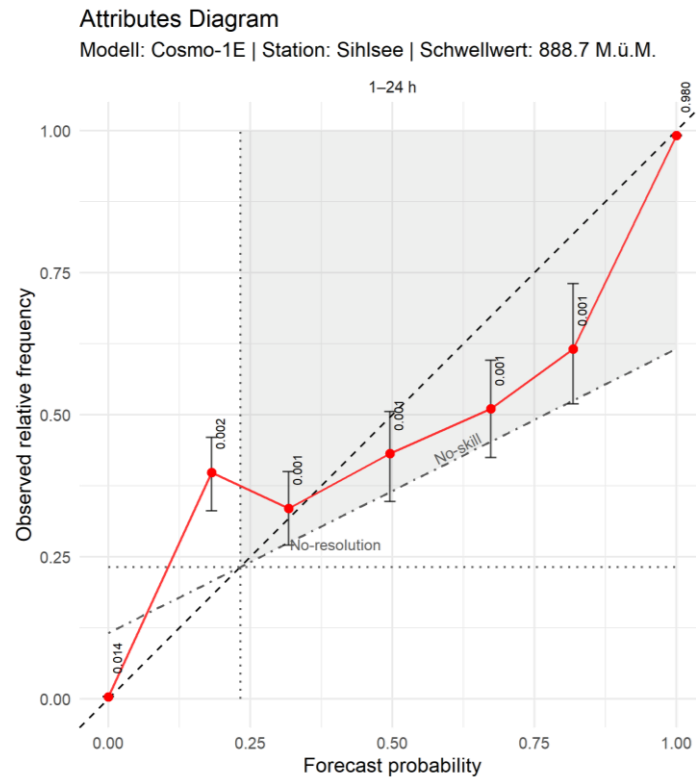


Abbildung 41: Attribut Diagramm des meteorologischen Modells Cosmo-1E für den Schwellwert 888.7 m. ü. M. und verschiedene Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2020-2024.

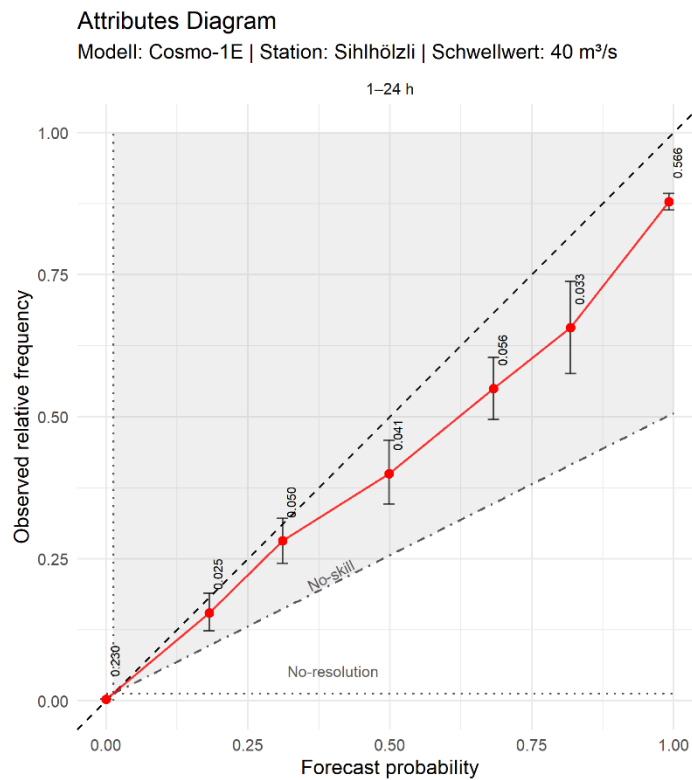


Abbildung 42: Attribut Diagramm des meteorologischen Modells Cosmo-1E für die Abflussschwelle von 40 m³/s und verschiedene Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2020-2024.

Attributes Diagram

Modell: ICON-1 | Station: Sihlsee | Schwellwert: 888.7 M.ü.M.

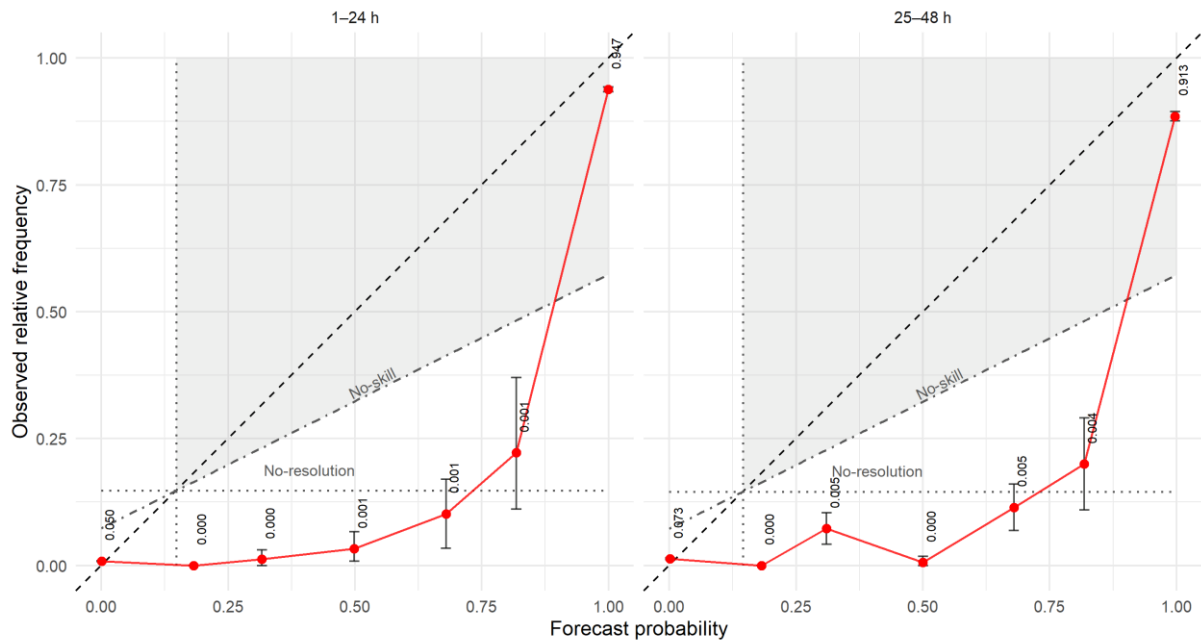


Abbildung 43: Attribut Diagramm des meteorologischen Modells Icon-1 für den Schwellwert 888.7 m. ü. M. und verschiedene Vorlaufzeiten für die Station Sihlsee im Zeitraum Mai bis November 2024-2025.

Attributes Diagram

Modell: ICON-1 | Station: Sihlhölzli | Schwellwert: 40 m³/s

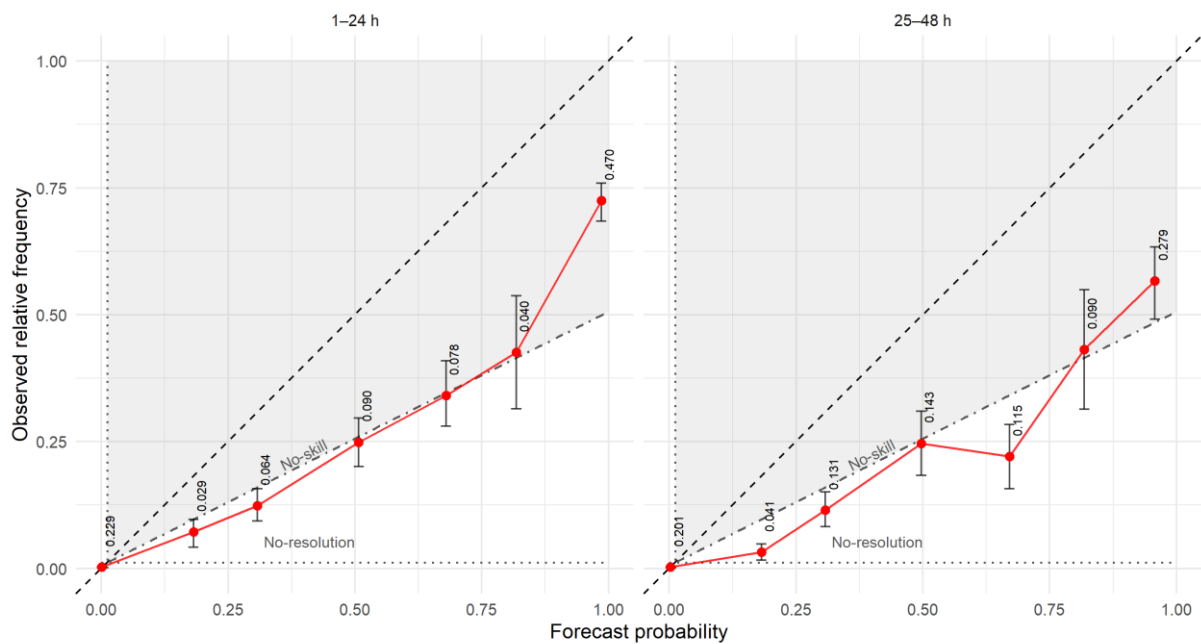
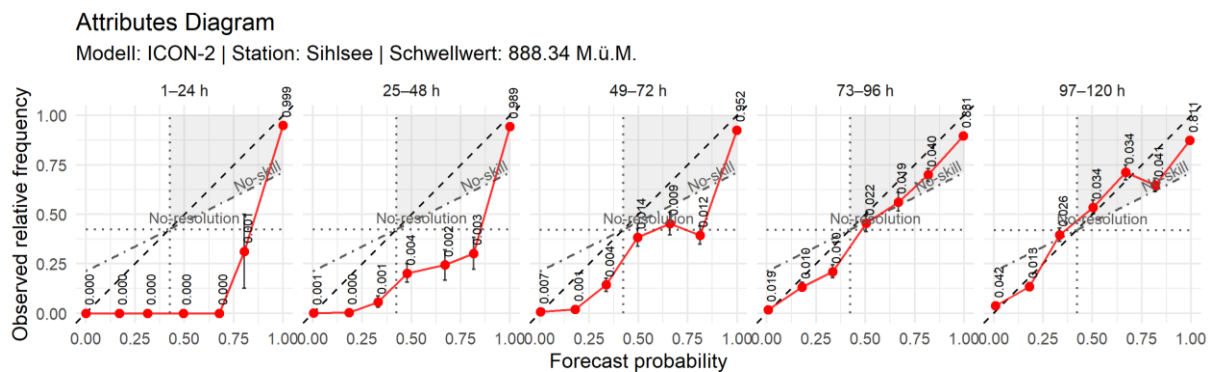
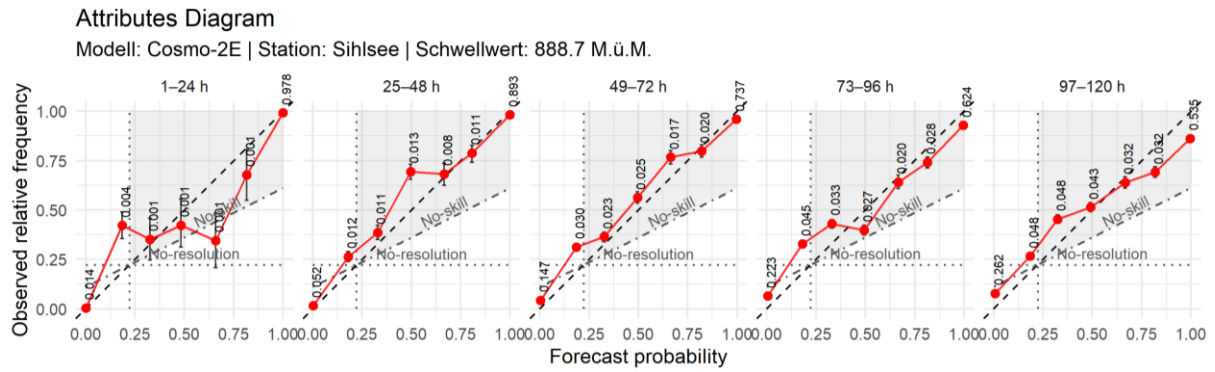


Abbildung 44: Attribut Diagramm des meteorologischen Modells Icon-1 für die Abflussschwelle von 40 m³/s und verschiedene Vorlaufzeiten für die Station Sihlhölzli im Zeitraum Mai bis November 2024-2025.



9.5 Eidesstattliche Erklärung

Persönliche Erklärung: Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

A handwritten signature in black ink, appearing to read 'A. Manente', with a stylized, cursive script.

Arianna Manente