Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

LANDESSTIFTUNG
Baden-Württemberg
Wir stiften Zukunft

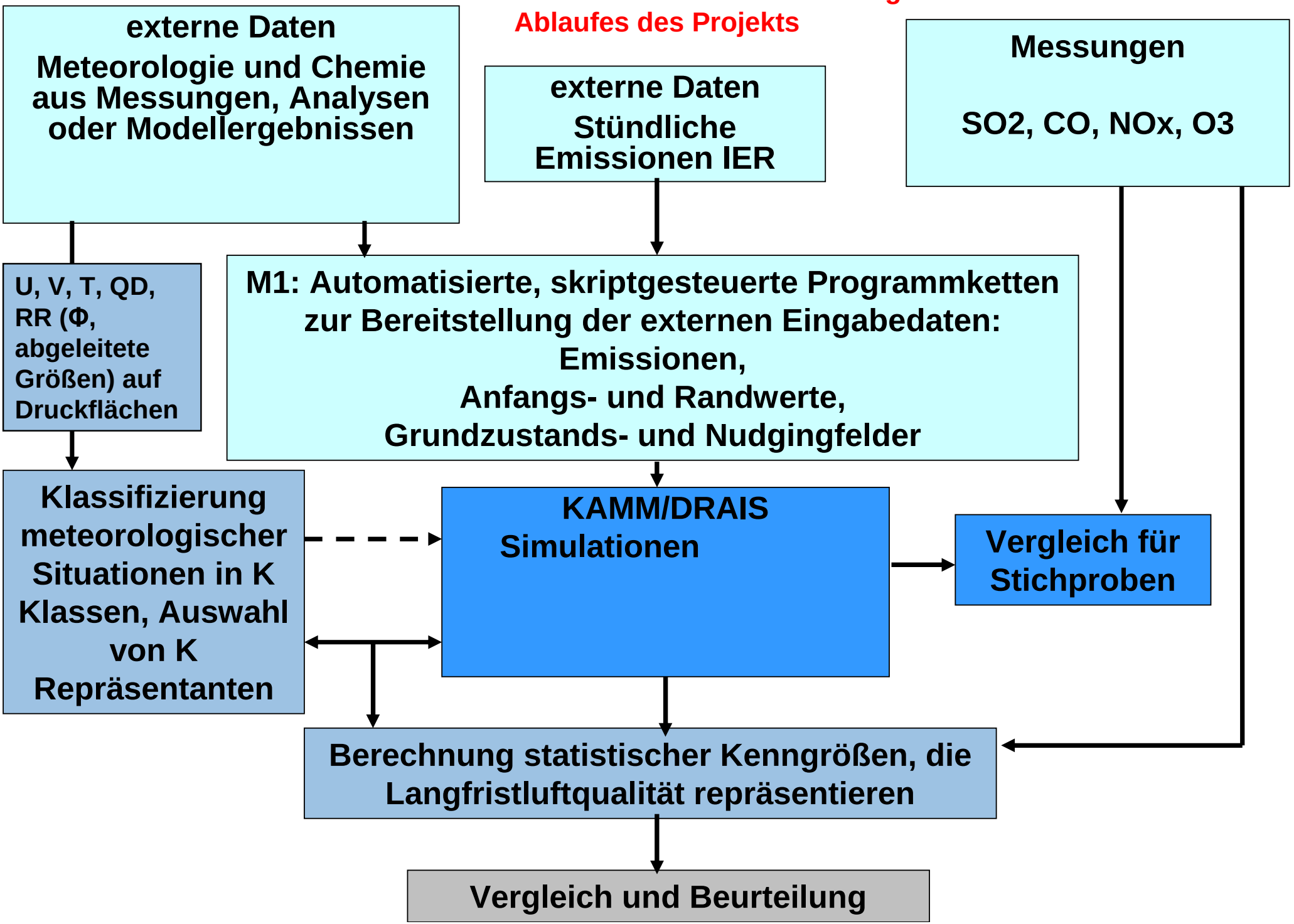
Optimierung der Modellierung der langfristigen Luftqualität

Hans-Jürgen Panitz
Forschungszentrum Karlsruhe
Institut für Meteorologie und Klimaforschung (IMK-TRO)

Projekt 740 finanziert durch die Landesstiftung Baden-Württemberg
Im Rahmen des Forschungsprogramms
„Modellierung und Simulation auf Hochleistungscomputern“
Projekt: 01.09.2004 – 31.08.2006

Zielsetzung und Vorgehensweise

Motivation und Schematische Darstellung des Ablaufes des Projekts



Ziel:

Bereitstellung einer Methode für die Klassifizierung meteorologischer Situationen in Hinblick auf ihre praktische Anwendbarkeit bei der Analyse der langfristigen Luftqualität einer Region

Untersuchung und Beurteilung von zwei verschiedenen Methoden:

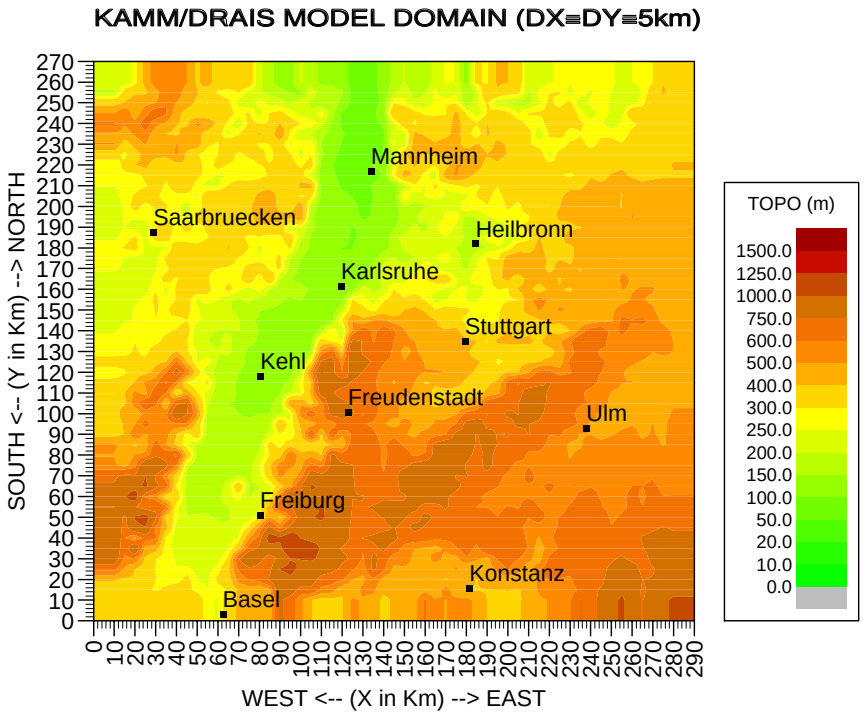
1. **„klassische" Clusteranalyse:**
   Ward Methode (hierarchisch agglomerativ)
   K-Means (partitionierender Algorithmus)

2. **SOM Technik (Self Organizing Maps)**
   vom Ansatz partitionierend, Methode der Zuordnung von Objekten und Klassen anders als z.B. K-Means

Beurteilungskriterien:

- **Effektivität der Methode (Handhabung, Rechenzeit, Interpretation der Klassifizierungsergebnisse)**
- **Repräsentieren die gefundenen Klassen das Spektrum der möglichen meteorologischen Bedingungen?**
- **Stimmen Luftqualitätsindikatoren, abgeleitet aus den Ergebnissen der Klassifizierung überein mit denen, die aus einer Detailberechnung von Schadstoffverteilungen berechnet wurden?**

KAMM/DRAIS Simulationen



NX =59 NY=55 NZ=35

Obergrenze: 5000m über NN

ΔX=5km ΔY=5km $\Delta Z_{Boden} \approx$10m $\Delta Z_{Top} \approx$ 250m

Start der Simulation: 1 Januar 00:00 UTC,

Ergebnisspeicherung: Δt = 1h

Numerischer Zeitschritt: ≤ 20 sec

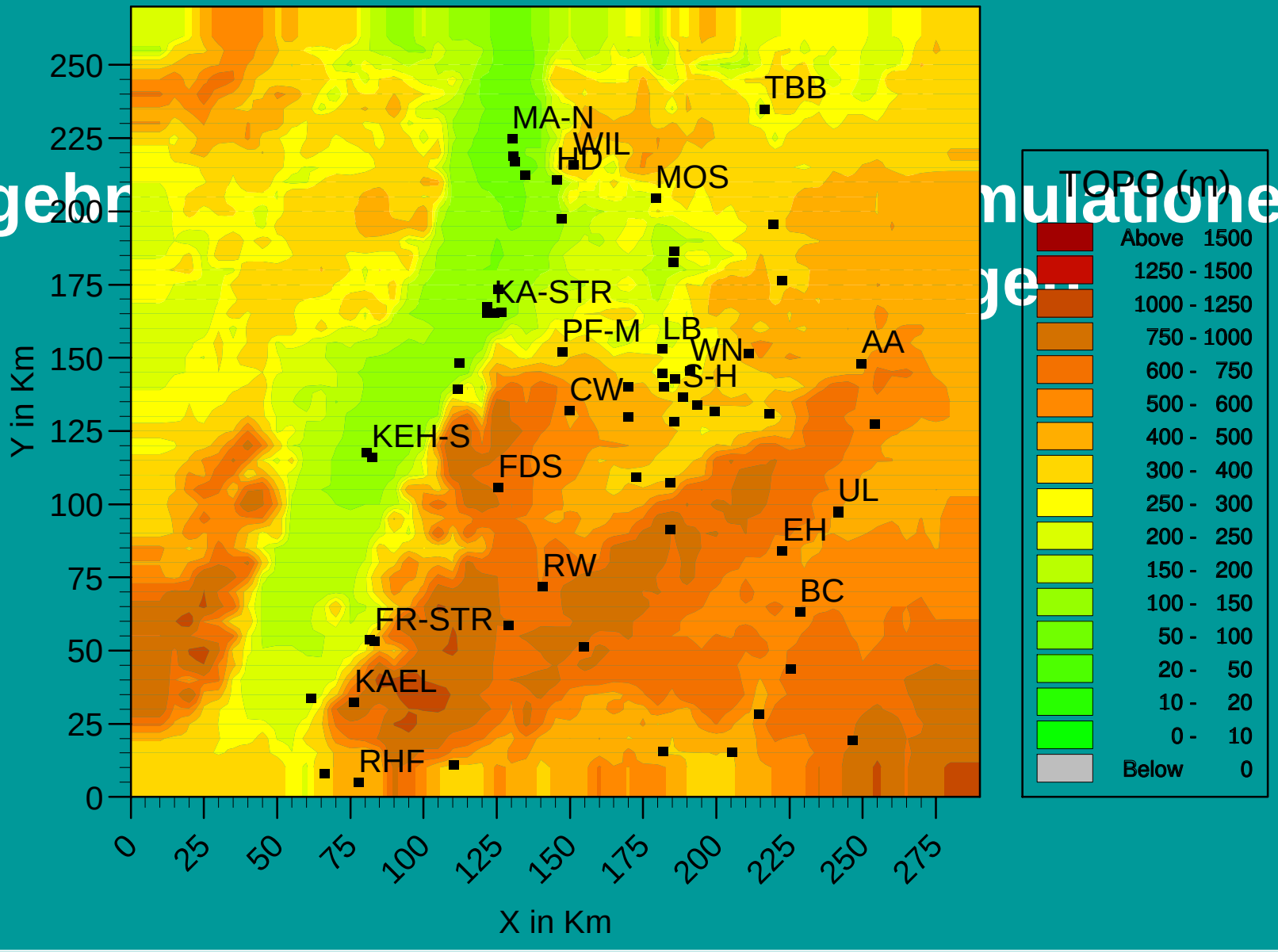
Anfangs- und Randbedingungen aus EURAD Ergebnissen:

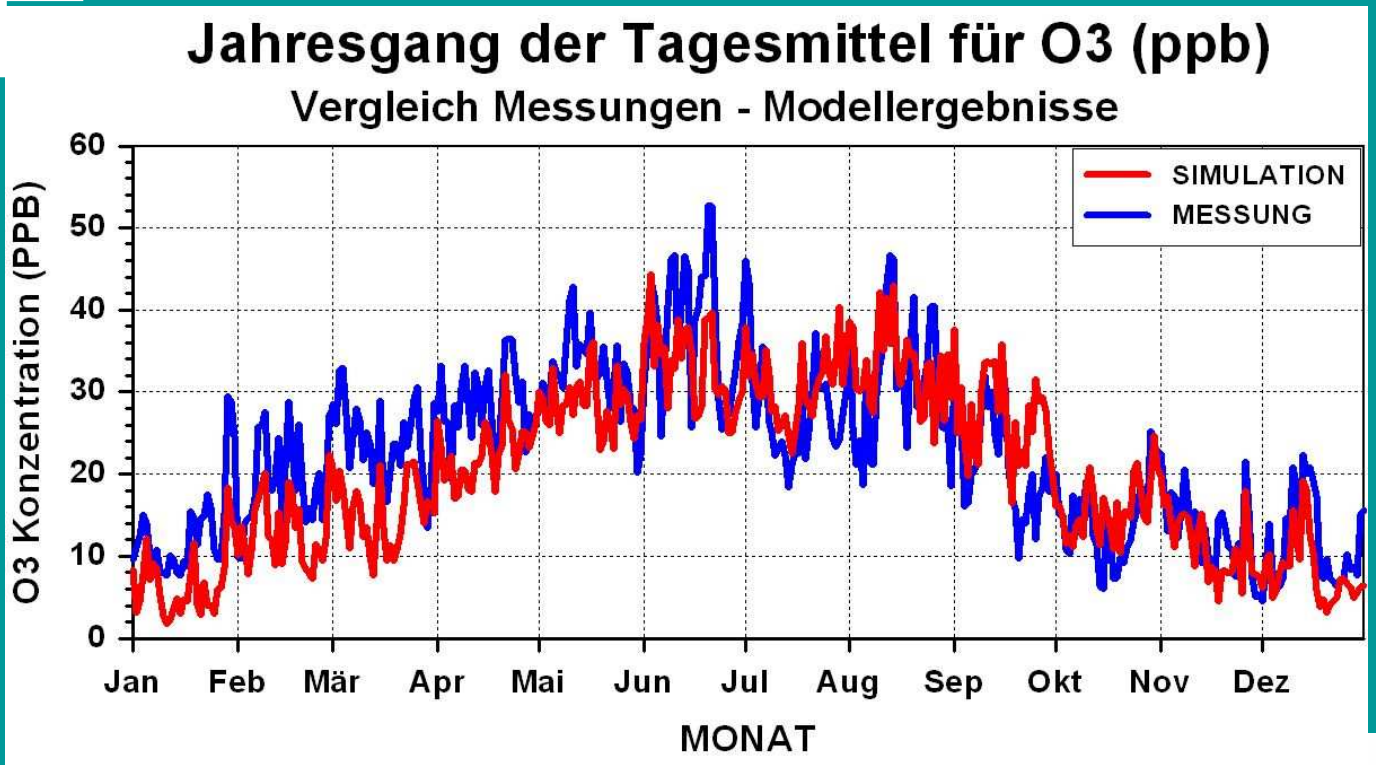
- **„Reinitialisierung“ jeden 2. Tag mit Einschwingphase von 3 h**
- **Randwerte werden stündlich neu eingelesen:**
  - **T, QD, and C durch Flussmethode: advektiver Transport über den Rand**
  - **Für u und v Orlanski Strahlungsbedingungen zu jedem numerischen Zeitschritt**

Großskaliger Grundzustand und Nudging Felder für u und v alle 3h neu, dazwischen lineare Interpolation

Nudging Koeffizient: 3.0E-4 konstant

Ergebn… …mulationen …ge…





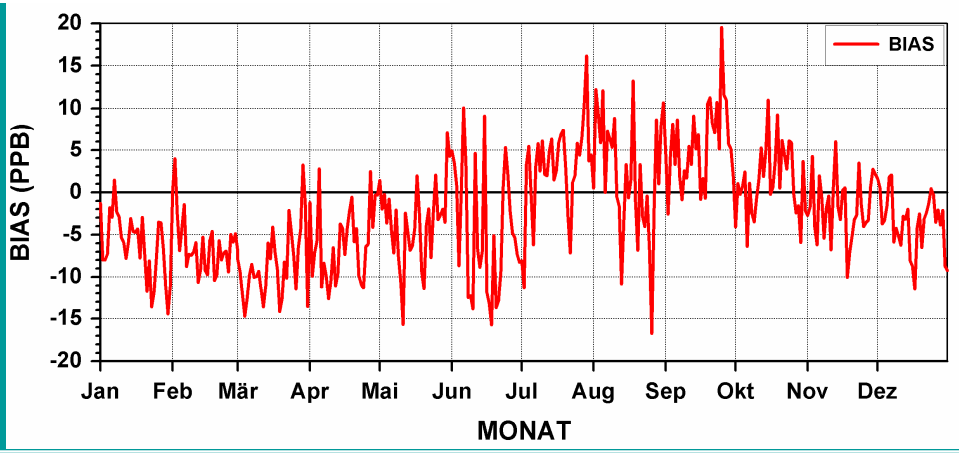
$$\text{Mittlerer Bias} = \frac{1}{N}\sum_{i}^{N}\left(S_i - O_i\right)$$

$$\text{MNB} = \frac{1}{N}\sum_{i}^{N}\frac{S_i - O_i}{S_i + O_i}$$

S_i = Modellergebnisse

O_i = Messwerte

Mittelwerte:
Simulation: 21.3 ppb
Messung: 23.9 ppb
Mittlerer BIAS: -2.6 ppb
MNB: -10.6 %
N: 427246 stündliche Werte



Vergleich Modellergebnisse mit Messdaten

- **Die Ergebnisse der Detailsimulation sind mit Blick auf das eigentliche Ziel des Projekts und unter Berücksichtigung von Bedingungen wie Auflösung des Rechengitters und Unsicherheiten in den Eingangsdaten zufrieden stellend.**
- **Auffallend beim Ozon: systematische Unterschätzung der gemessenen Werte in der ersten Jahreshälfte**

Mögliche Gründe für Diskrepanzen:

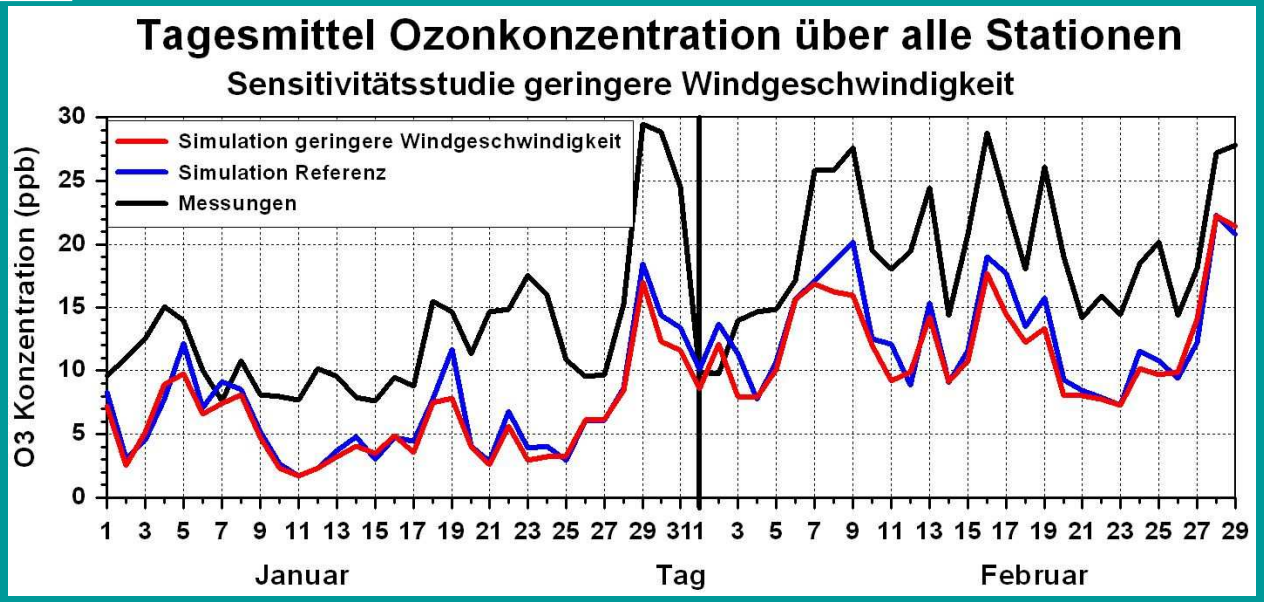
- **Auflösung des Modellgitters zu grob. Punktmessungen werden mit Flächenmittelwerten verglichen;**
- **Parametrisierung der Ozonchemie für den Winter und das Frühjahr hat Defizite;**
- **anderen „Ozonquellen“, z.B. Intrusion aus der Stratosphäre;**
- **Ergebnisse der meteorologischen Simulationen zu ungenau;**
- **Anfangs- und Randbedingungen stimmen nicht;**
- **Emissionsdaten sind nicht genau genug**

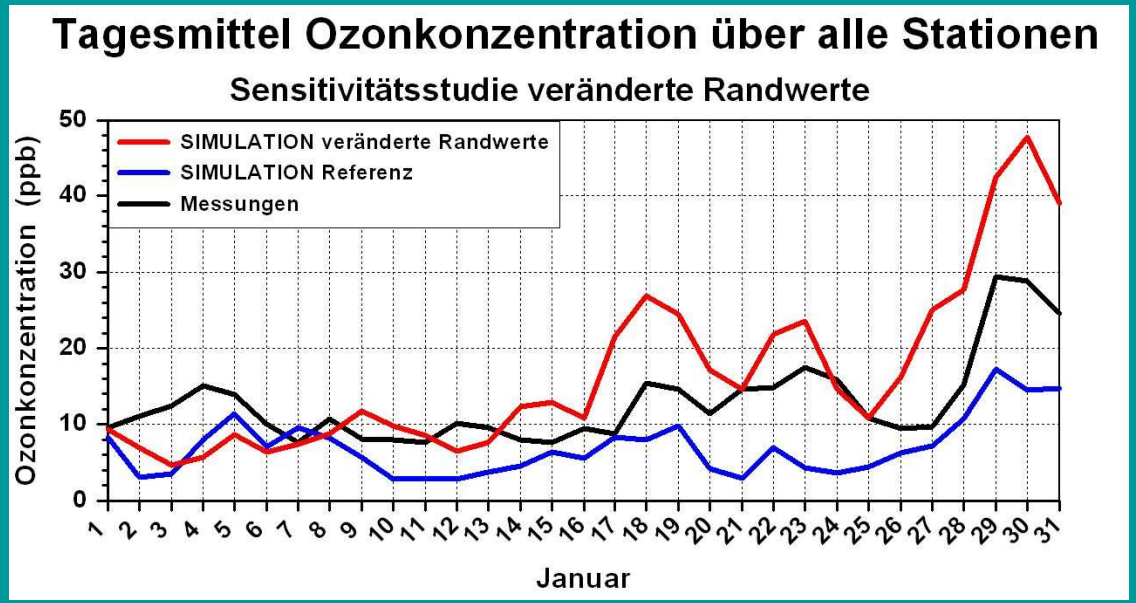
Sensitivitätsstudie für Monat Januar

	Mittelwert (ppb) (Messwert)	Kein Regen	Verdoppelung Niederschlagsintensität	Geringere Wind-geschwindigkeit	Anfangs - werte	Ränder
O_3	6.7 (13.4)	Leichte Abnahme (-4.3%)	Leichte Zunahme (1.2%)	Abnahme (-9.7%)	Zunahme (5.5%)	Zunahme (154.5%)
NO_X	36.0 (39.6)	Leichte Zunahme (1%)	Leichte Abnahme (-0.7%)	Zunahme (19.4%)	Abnahme (-2.1%)	Abnahme (-6.6%)

Größter Einfluss durch Veränderung

- **der Windgeschwindigkeit, aber falsche Richtung, Ozon nimmt weiter ab**
- **der Randwerte, Erhöhung der Ozonwerte unrealistisch**


Tagesmittel Ozonkonzentration über alle Stationen
Sensitivitätsstudie geringere Windgeschwindigkeit
Simulation geringere Windgeschwindigkeit
Simulation Referenz
Messungen
O3 Konzentration (ppb)
Januar
Tag
Februar


Tagesmittel Ozonkonzentration über alle Stationen
Sensitivitätsstudie veränderte Randwerte
SIMULATION veränderte Randwerte
SIMULATION Referenz
Messungen
Ozonkonzentration (ppb)
Januar

Clusteranalyse

Verwendete Methoden

- **Klassische Clusteranalyse: WARD, K-MEANS**
- **SOM Technik**

Klassifizierte Objekte:

- **tägliche meteorologische Bedingungen des Jahres 2000**
- **Charakterisiert durch Tagesmittelwerte der verschiedener meteorologischen Größen**
- **Alle Größen ermittelt aus Ergebnissen des EURAD Modells an 11*10 = 110 Gitterpunkten**
- **2 Parameterkombinationen:**

Kombination	bodennah	850 hPa	700 hPa	500 hPa	# Variable
1	Ps, U10, V10	U, V, T, Φ	U, V, T, Φ	U, V, T, Φ	1650 = 15*110
2	Ps, U10, V10, T2, Qd2	U, V, T, Φ, Qd	-	U, V, T, Φ, Qd	1650 = 15*110

Größe der Datenmatrizen sowie Zahl der signifikanten EOFs und die jeweiligen erklärten Anteile an der Gesamtvarianz für verschiedene Klassifikationszeiträume

Parameterkombination 1				
Zeitspanne	Größe Rohdatenmatrix	Anzahl signifikanter EOFs	Erklärter Varianzanteil (%)	Größe Ausgangsmatrix für Clusteranalyse
Jahr	366 x 1650	12	95.6	366 x 12
April-September	183 x 1650	10	90.0	183 x 10
Mai-Juli	92 x 1650	7	90.2	92 x 7

Parameterkombination 2				
Zeitspanne	Größe Rohdatenmatrix	Anzahl signifikanter EOFs	Erklärter Varianzanteil (%)	Größe Ausgangsmatrix für Clusteranalyse
Jahr	366 x 1650	14	93.2	366 x 14
April-September	183 x 1650	12	89.8	183 x 12
Mai-Juli	92 x 1650	9	87.5	92 x 9

Problem einer jeden Klassifizierung

Bestimmung der „optimalen“ Anzahl von Klassen

- in Anlehnung an Ausreißeranalyse von Bacher (2002), Definition eines Homogenitäts-/Ausreißerkriteriums:
- Eine Clusterlösung mit K Klassen ist homogen/ausreißerfrei falls

$$d_j = \sqrt{\frac{d^2_{jk}}{d^2_0}} \leq 1$$

d_j = standardisierte Distanz des Objektes j vom Zentrum des Clusters g_k

d^2_0 = mittlere quadrierte euklidischen Distanz aller Objekte vom Zentrum der 1-Clusterlösung

d^2_{jk} = Summe über die quadrierten euklidischen Distanzen der Variablen i innerhalb einer Klasse g_k vom jeweiligen Clusterzentroid $\overline{x}_{g_k i}$

Bacher, J., 2002: Clusteranalyse, Oldenbourg Verlag

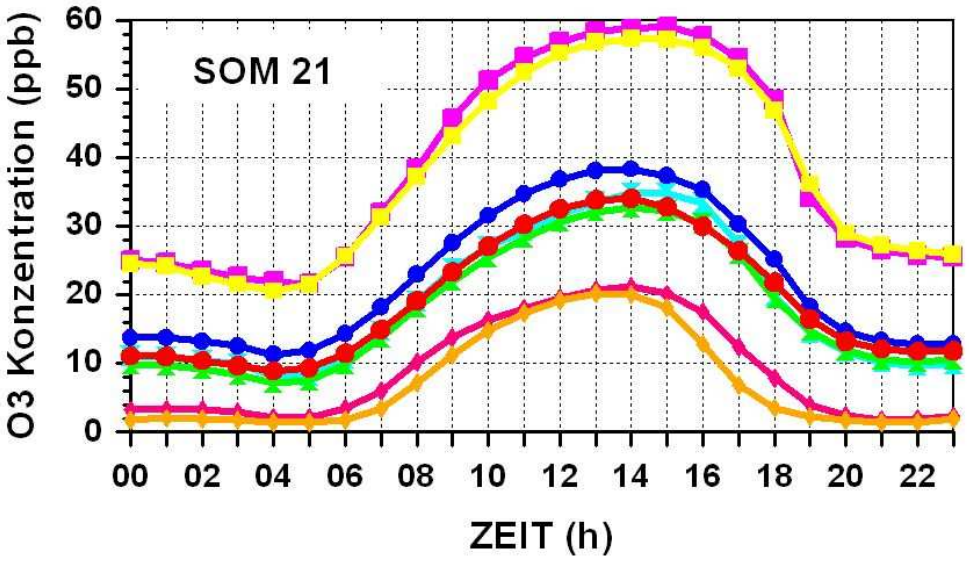
Ergebnisse Clusteranalyse

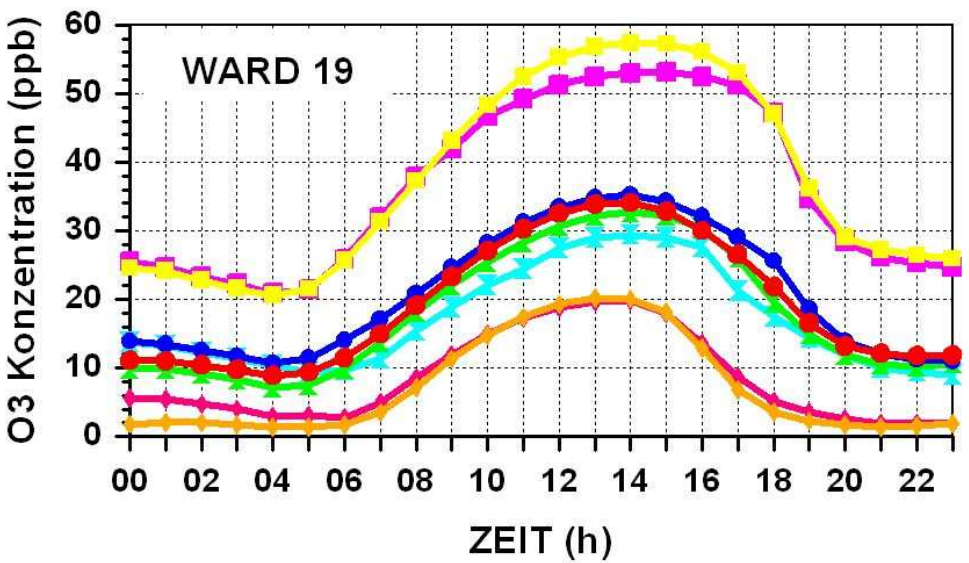
Parameterkombination 1

Methode	Rechenzeit (sec)	NCMIN	NCMAX	Clusterzahl
SOM	700	1	100	**21**
WARD	160	1	365	**19**
K-Means	200	2	365	**20**

Ergebnisse Clusteranalyse

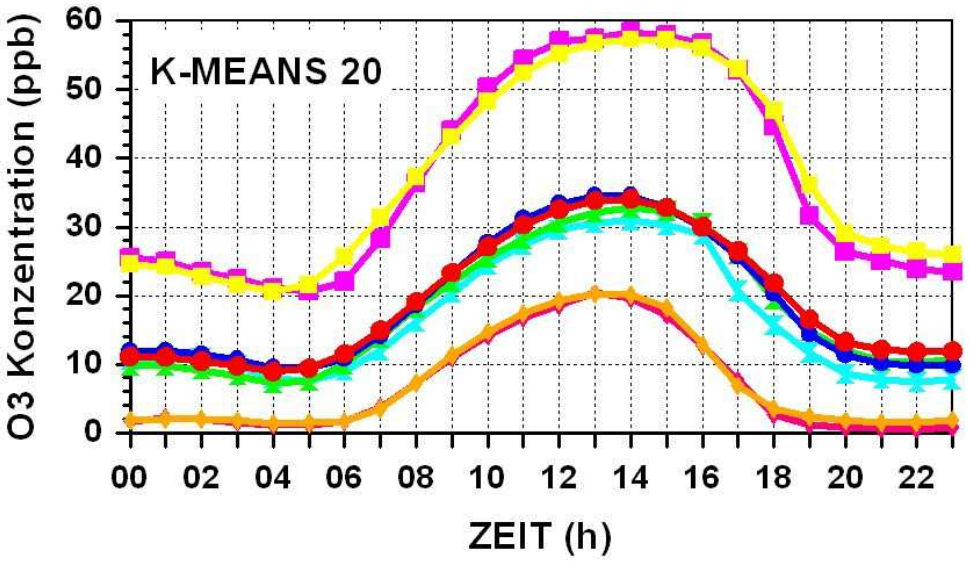
Vergleich Luftqualitätsindikatoren resultierend aus Detailsimulation bzw. Clusteranalysen

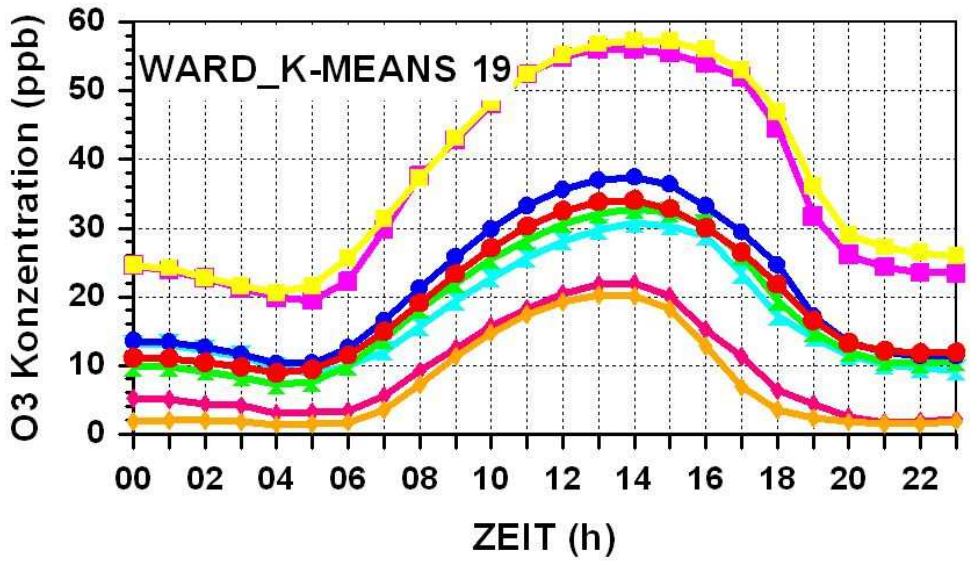




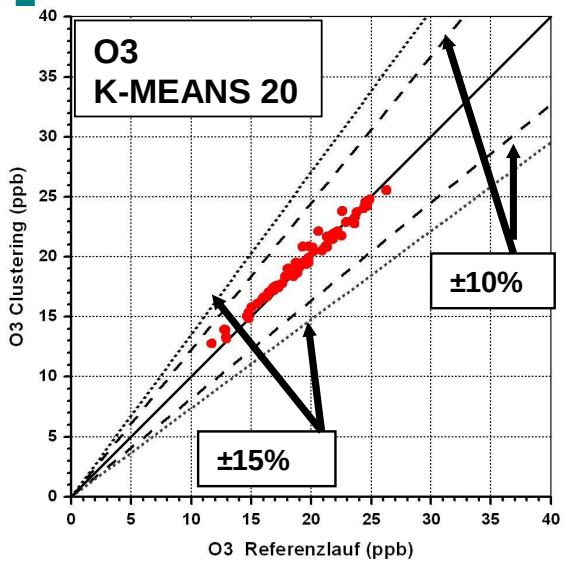
SIMULATION
CLUSTERING
REFERENZ Median
CLUSTERING Median
REFERENZ 25-Perzentil
CLUSTERING 25-Perzentil
REFERENZ 90-Perzentil
CLUSTERING 90-Perzentil

Abweichungen < 7%

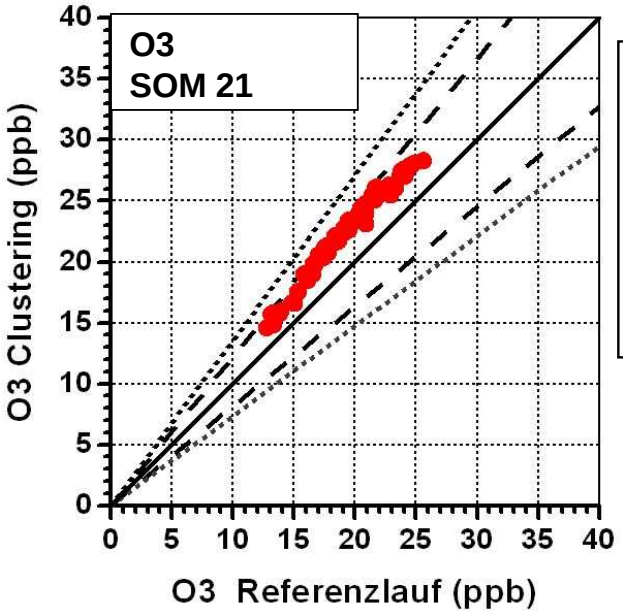




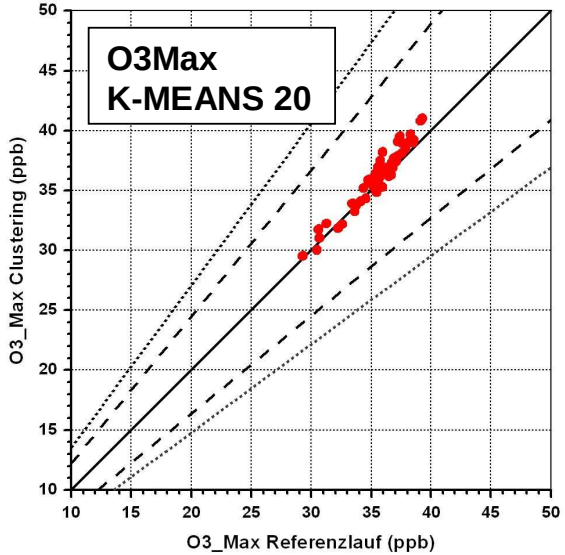
Mittlere Tagesgänge Ozon. Mittelung über alle Stationen



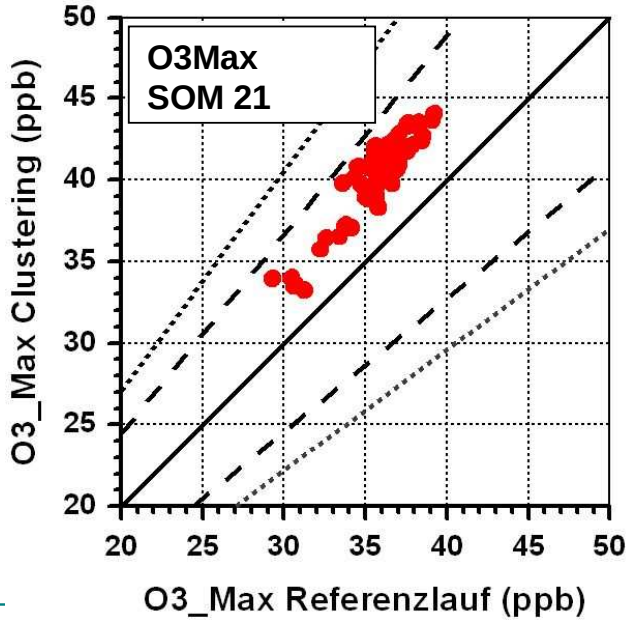
Klassifikation:
62 Stationen
Mittelwerte O3:
Referenz: 19.3 ppb
K-MEANS 20: 19.1 ppb
Bias: - 0.2 ppb
Rel. Bias - 0.7 %



Klassifikation:
62 Stationen
Mittelwerte O3:
Referenz: 19.3 ppb
SOM 21: 22.5 ppb
Bias: 3.2 ppb
Rel. Bias 7.6 %

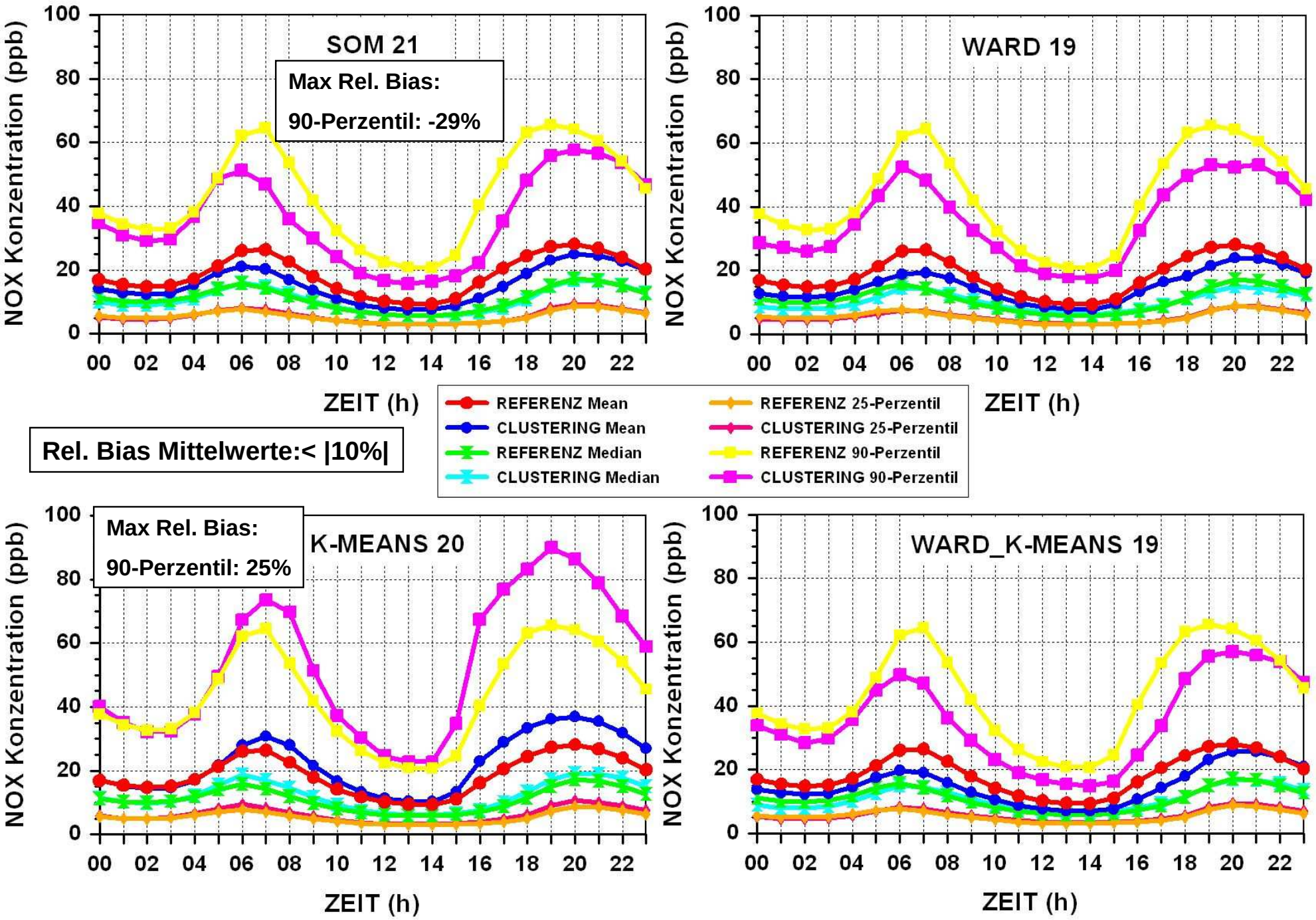


Klassifikation:
62 Stationen
Mittelwerte O3Max:
Referenz: 35.6 ppb
K-MEANS 20: 36.3 ppb
Bias: 0.7 ppb
Rel. Bias 0.9 %



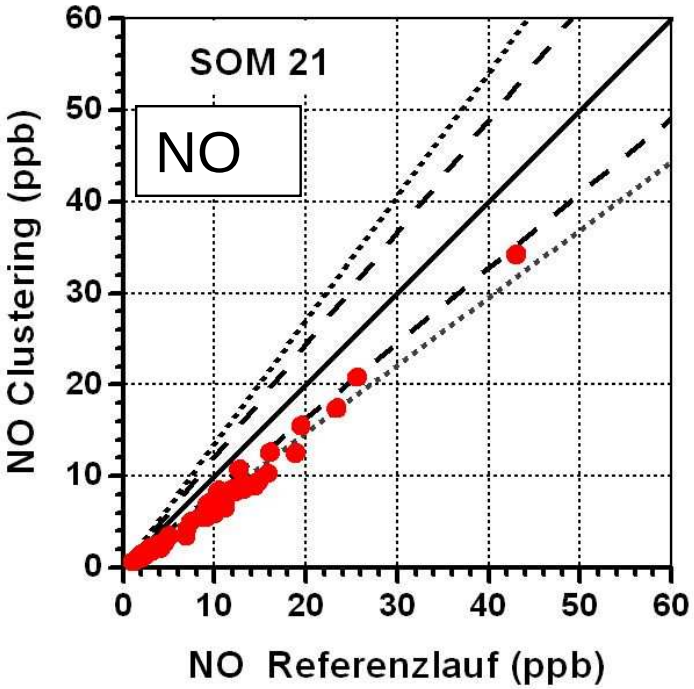
Klassifikation:
62 Stationen
Mittelwerte O3Max:
Referenz: 35.6 ppb
SOM 21: 40.2 ppb
Bias: 4.6 ppb
Rel. Bias 6.0%

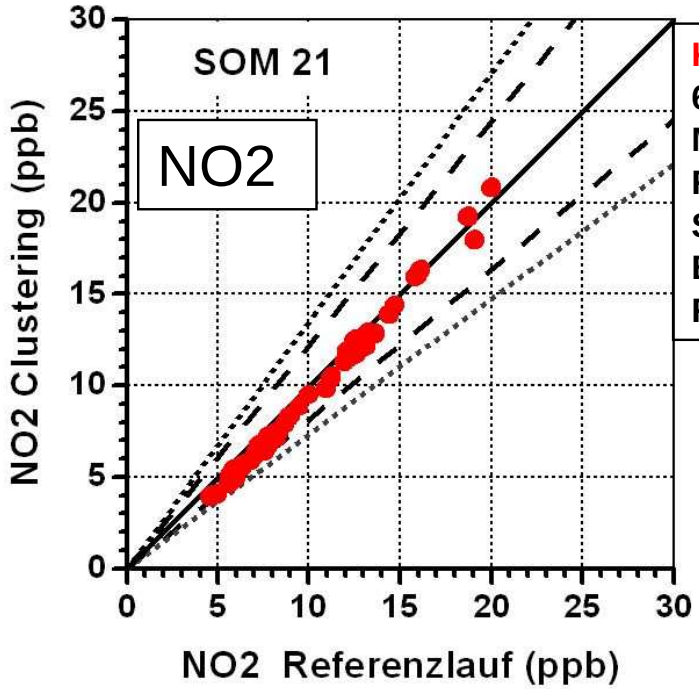
Vergleich Jahresmittelwerte an Messstationen



Mittlere Tagesgänge NOx. Mittelung über alle Stationen

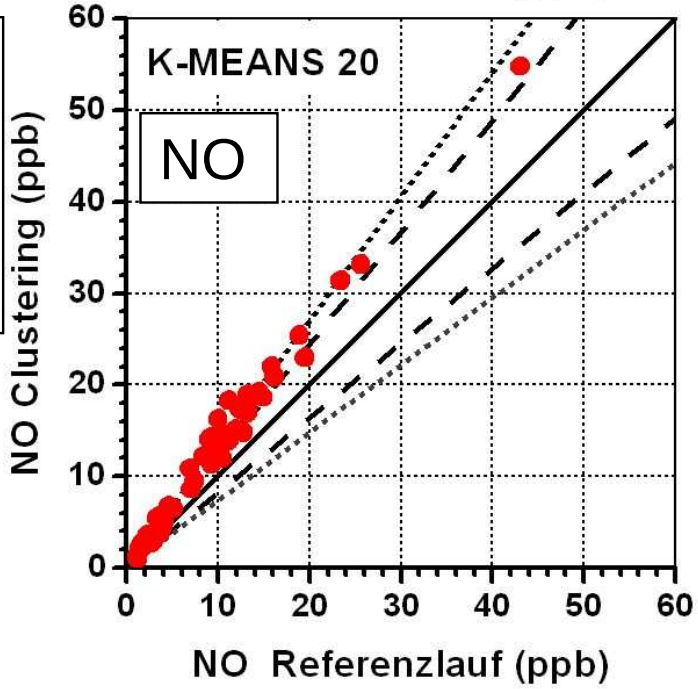
Klassifikation:
62 Stationen
Mittelwerte NO:
Referenz: 8.3 ppb
SOM 21: 5.7 ppb
Bias: - 2.6 ppb
Rel. Bias: -18.3 %

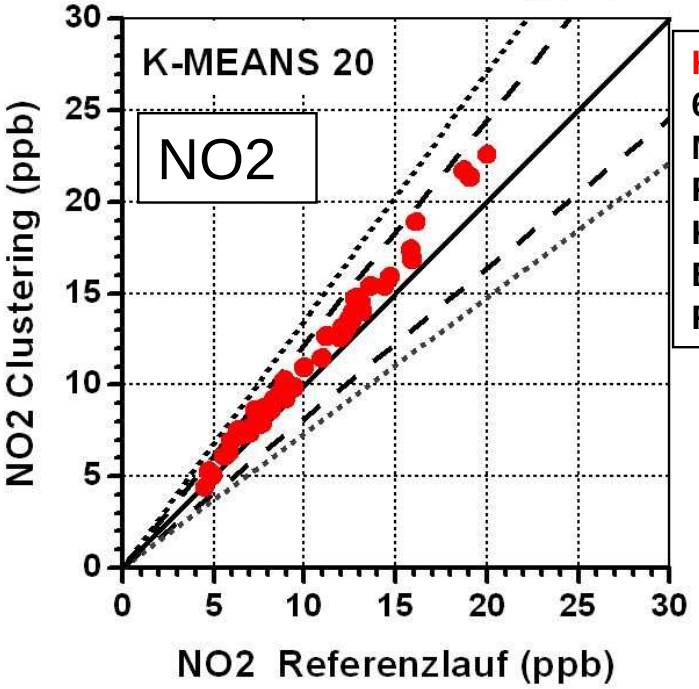




Klassifikation:
62 Stationen
Mittelwerte NO2:
Referenz: 10.4 ppb
SOM 21: 9.8 ppb
Bias: -0.6 ppb
Rel. Bias: -3.0 %

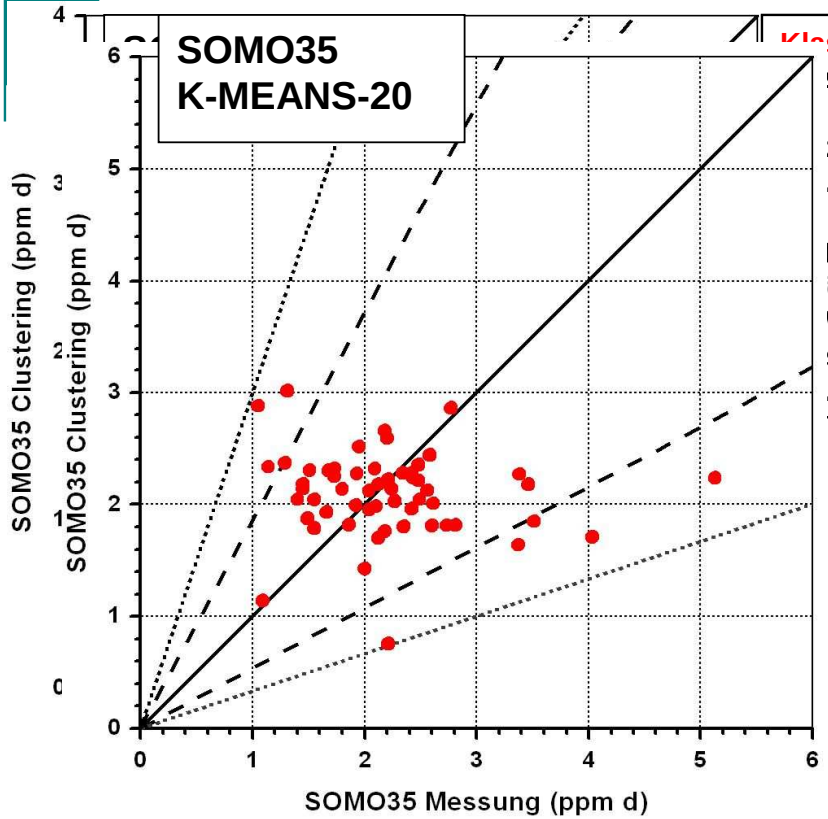
Klassifikation:
62 Stationen
Mittelwerte NO:
Referenz: 8.3 ppb
K-MEANS: 11.0 ppb
Bias: - 2.7 ppb
Rel. Bias: 14.1%





Klassifikation:
62 Stationen
Mittelwerte NO2:
Referenz: 10.4 ppb
K-MEANS: 11.4 ppb
Bias: 1.0 ppb
Rel. Bias: 4.6 %

Vergleich Jahresmittelwerte an Messstationen

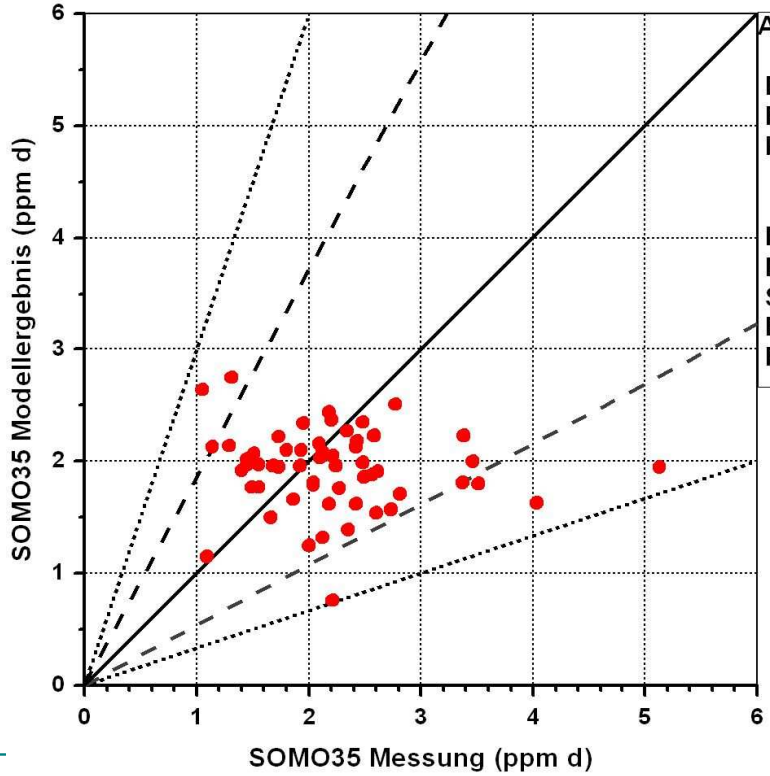


Klassifikation
Anzahl Stationen: 62

Einzelwerte:
Rel. Bias ≤ |10%|; 62 (100%)

Mittelwerte:
Differenz: 2.0 ppm d
K-Means-20: 2.1 ppm d
Bias: 0.1 ppm d
Rel. Bias: 4.4 %

Vollständige Simulation



Anzahl Stationen: 56

Einzelwerte:
Rel. Bias <= |30%|: 48 (86%)
Rel. Bias <= |50%|: 56 (100%)

Mittelwerte:
Messwerte: 2.2 ppm d
Simulation: 1.9 ppm d
Bias: -0.3 ppm d
Rel. Bias: -4.9%

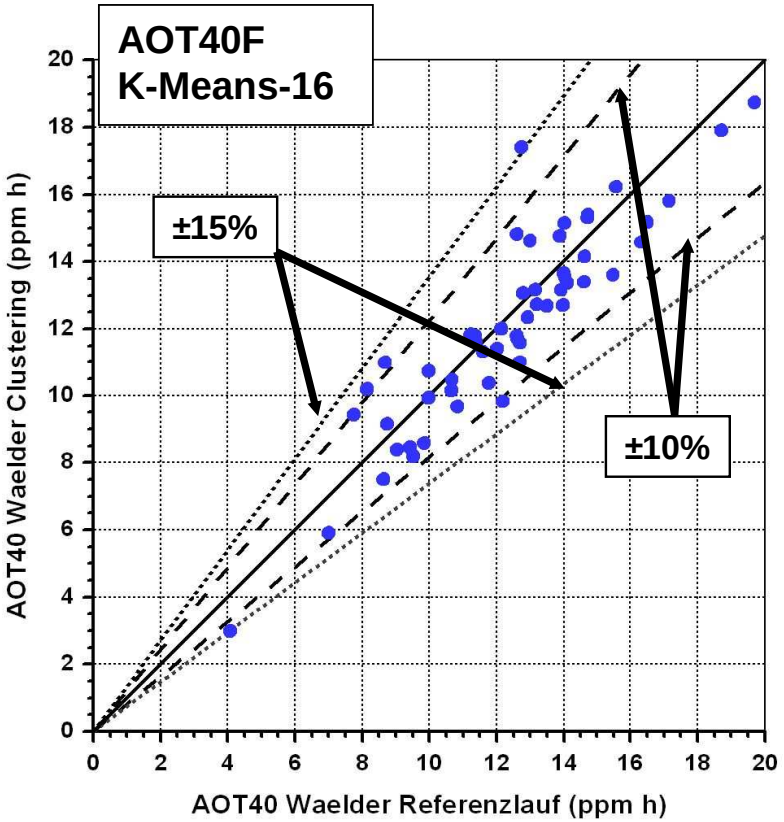
SOMO35: (Sum of Ozone Means over 35 ppb)

Jährliche Summe über die täglichen Maxima der 8-stündigen gleitenden Mittelwerte, die 35 ppb überschreiten:

$$SOMO35 = \sum_{d=1}^{Ny} \max\left(A_8^d - 35ppb, 0\right)$$

AOT40: (accumulated amount of ozone above 40 ppb)
Integrationsintervall AOT40 für Wälder:
April-September, 08:00 – 20:00 Lokalzeit

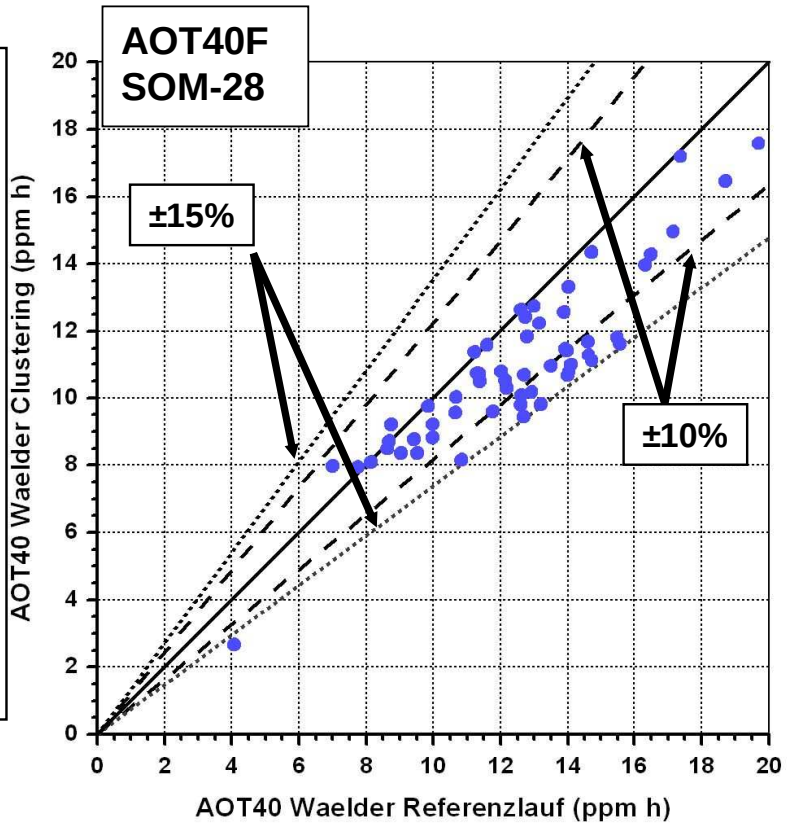
$$AOT\ 40 = \int \max\left(O_3 - 40\ \text{ppb}\ ,0\right) dt$$



Klassifikation
62 Stationen

Mittelwerte K-Means
Referenz: 12.6 ppm h
K-Means16: 12.4 ppm h
Bias: -0.2 ppm h
Rel. Bias: -1.1%

Mittelwerte SOM
Referenz: 12.6 ppm h
SOM-28: 11.1 ppm h
Bias: -1.5 ppm h
Rel. Bias: -6.3%



Bewertung

System mit Methoden zur Klassifizierung meteorologischer Bedingungen mit Hinblick auf Anwendungen zur Bestimmung der langfristigen Luftqualität in Baden-Württemberg wurde entwickelt

Statistische Kenngrößen, die die Luftqualität beschreiben, lassen sich mit genügender Genauigkeit auf Basis der Ergebnisse der Klassifizierung berechnen

Rückschluss auf tatsächliche langfristige Luftqualität in B.-W. nicht möglich, dazu muss Zeitraum von 10-20 Jahren betrachtet werden

Bewertung

Hinsichtlich der Kriterien

Effizienz der Methode und Repräsentativität der gefundenen Klassen

erscheint K-Means Verfahren am besten geeignet (im Mittel geringste Abweichungen zu den Ergebnissen aus vollständiger Jahressimulation)

Ergebnisse basierend auf SOM weisen im Mittel größte Abweichungen auf; außer Notwendigkeit zur Festlegung der Zahl der Klassen weitere Parametervorgaben nötig, die SOM beeinflussen

Aspekte und Fragestellungen, die sich aus dem Projekt ergeben haben

1. **Sensitivitätsstudie mit unterschiedlichen Parametereinstellungen bei SOM**
2. **Luftqualität nicht nur abhängig von den meteorologischen Bedingungen, sondern auch von Emissionen, die nicht nur von der Tageszeit, sondern auch vom Wochentag (Werktag/Wochenende) und von den Jahreszeiten abhängen. Um noch bessere Abschätzungen der Luftqualität zu erreichen**
   - **Müssen Emissionen mit in die Klassifizierung einbezogen werden?**
   - **Müssen für jede gefundene Klasse mittlere oder gewichtete Emissionsverteilungen sowie mittlere Antriebs- und Randdaten erstellt werden, die dann als Eingabe neuer Modellrechnungen verwendet werden?**
3. **Klassifizierung von realen, also nicht der von einem Modell berechneten meteorologischen Größen (direkte Messungen oder datenassimilierte Analysen)**
4. **Beurteilung der Ergebnisse der Klassifizierung aus meteorologischer Sicht. Repräsentativität der gefundenen Klassen wurde nur über die Qualität der Reproduktion der statistischen Luftqualitätskenngrößen beurteilt**

Aspekte und Fragestellungen, die sich aus dem Projekt ergeben haben

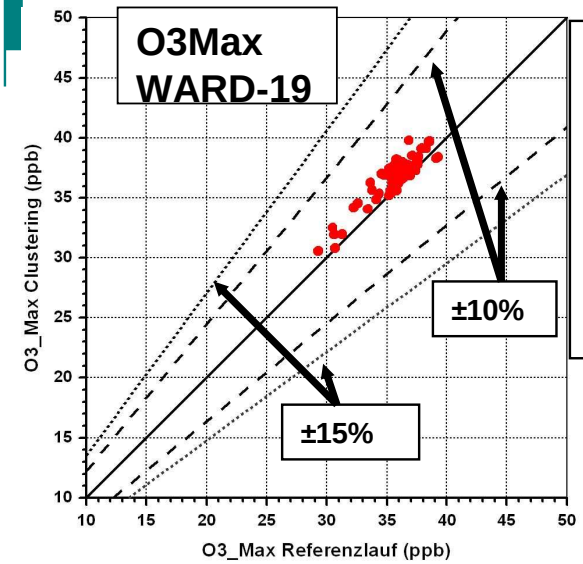
5. **Systematische Untersuchung der Gründe für die Unterschätzung der Ozonkonzentrationen durch Modell in der ersten Jahreshälfte, z.B. durch Anwendung eines zum Teil noch zu entwickelnden adjungierten KAMM/DRAIS Modells**
6. **Austausch des Modellsystems KAMM/DRAIS durch das System LM-ART (Vogel, B. und Mitarbeiter, 2006; ART: Aerosols and Reactive Trace gases), in dem das Ausbreitungsmodell DRAIS an das Lokal Modell LM des Deutschen Wetterdienstes angekoppelt wurde. Dadurch Erweiterung der Berechung der Luftqualitätsindikatoren auf Feinstäube und Pollen möglich.**

Vogel, B. und Mitarbeiter (2006): http://www-imk.fzk.de/tro/ACP/index.html

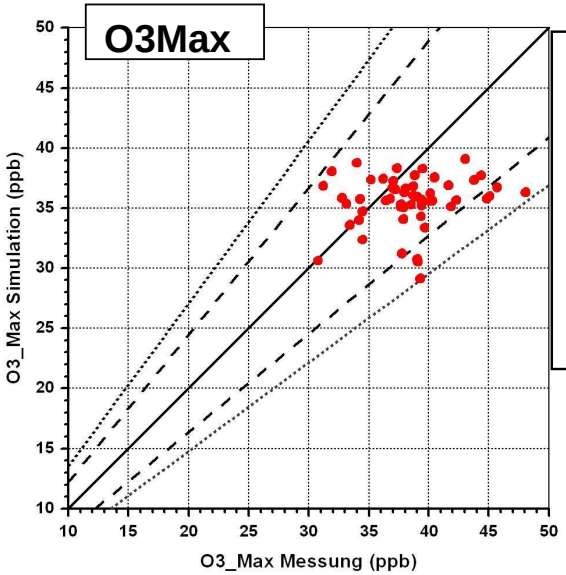
Danksagung

Ich bedanke mich

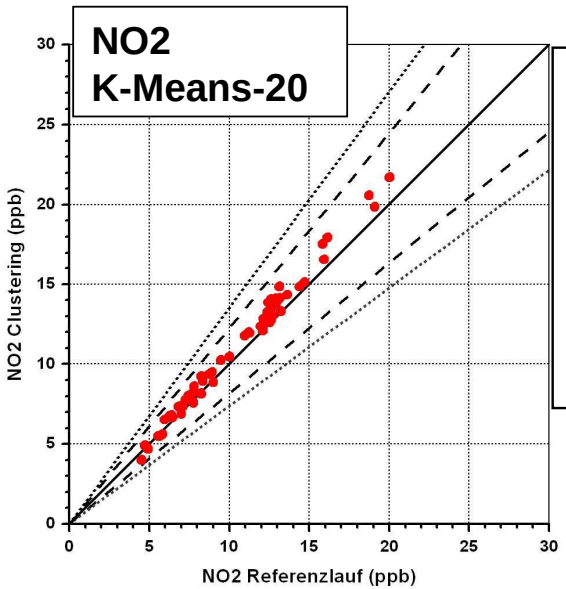
- **bei der Landesstiftung Baden-Württemberg für die Finanzierung des Projektes**
- **beim Zentrum für Umweltmessungen, Umwelterhebungen und Gerätesicherheit (UMEG) für die Bereitstellung der umfangreichen Messdaten**
- **bei den Kollegen des Instituts für Energiewirtschaft und Rationelle Energieanwendung (IER, Universität Stuttgart) und des Rheinischen Instituts für Umweltforschung (Universität zu Köln, EURAD Gruppe) für die zügige Berechnung und Übermittlung der für das Modellsystem KAMM/DRAIS notwendigen Eingangsdaten**
- **bei Herrn Dipl.-Met. Klaus Nester für die Hilfe bei der Erstellung des Präprozessorsystems und der Umstellung des Modellsystems KAMM/DRAIS für Langzeitsimulationen**
- **bei Prof. Dr. Johann Bacher, Johann Kepler Universität Linz, für die umfangreiche Statistiksoftware, die auch Programme zur klassischen Clusteranalyse enthält**
- **bei Herrn Dr. Gerd Schädler für die Hilfe beim Umgang mit der SOM-Methode**



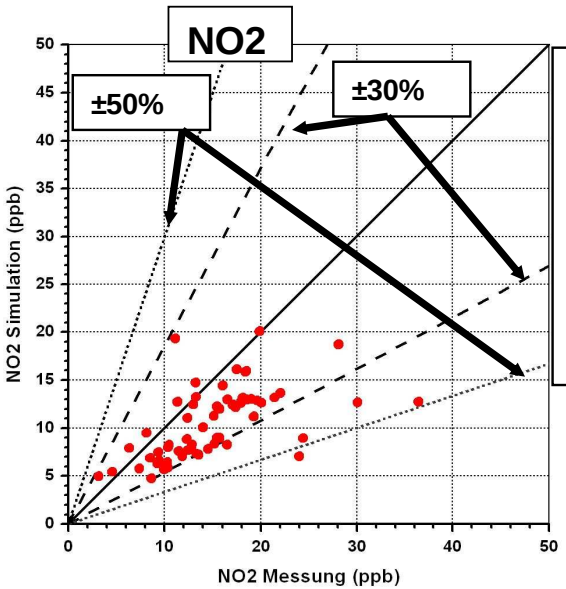
Klassifikation:
62 Stationen
Mittelwerte O3Max:
Referenz: 35.6 ppb
WARD-19: 36.7 ppb
Bias: 1.1 ppb
Rel. Bias 1.4%



Vollständige Simulation:
56 Stationen
Mittelwerte O3Max:
Simulation: 35.7 ppb
Messung: 38.7 ppb
Bias: 3.0 ppb
Rel. Bias -4.0%



Klassifikation:
62 Stationen
Mittelwerte NO2:
Referenz: 10.4 ppb
K-Means-20: 11.4 ppb
Bias: 1.0 ppb
Rel. Bias 4.6%



Vollständige Simulation:
62 Stationen
Mittelwerte NO2:
Simulation: 10.5 ppb
Messung: 15.1 ppb
Bias: -4.6 ppb
Rel. Bias -19.4%

Vergleich Jahresmittelwerte an Messstationen

Bezug zum Hochleistungsrechnen

- Notwendigkeit der Nutzung eines Hochleistungsrechners ergibt sich aus der Struktur des Modellsystems: 3-D Gitterpunktsmodell, konzipiert für meteorologische und luftchemische Untersuchungen über strukturiertem Gelände, numerische Lösung der Navier-Stokes-Gleichung und Bilanzgleichungen für 41 chemisch reaktive Spurengase

- Anwendung des bereits existierenden und für einen Vektorrechner konzipierten Modellsystems KAMM/DRAIS

- Keine Neuentwicklung jedoch Modifikationen, um der Simulationszeit von einem Jahr Rechnung zu tragen (Nasschemie, jahreszeitliche Veränderungen der Vegetation, I/O Strukturen)

- Verwendeter Rechner: VPP5000 (Vektor-Parallel-Rechner) des FZK
  - Max. 8 Processing Elements (PE). Jede PE hat Vector Unit mit 9.6 GFlop/s und Scalar Unit mit 1.2 GFlop/s, 6 PEs mit 8 Gbyte, 2 mit 16 Gbyte Hauptspeicher
  - Rechenzeit für gesamte Simulation für das Jahr 2000: ca. 1700 CPU Stunden ≈ 71 Tage

Bestimmung der Clusterzahl durch ein Homogenitätskriterium

Situation: n Objekte, charakterisiert durch N Variable, wurden in K Gruppen klassifiziert. Jedes Cluster besteht aus n_k Objekten, k = 1,…, K. Es gilt also $n = \sum_{k=1}^{K} n_k$.

Die Klassifizierung wird beendet, wenn alle Objekte j das Homogenitätskriterium erfüllen:

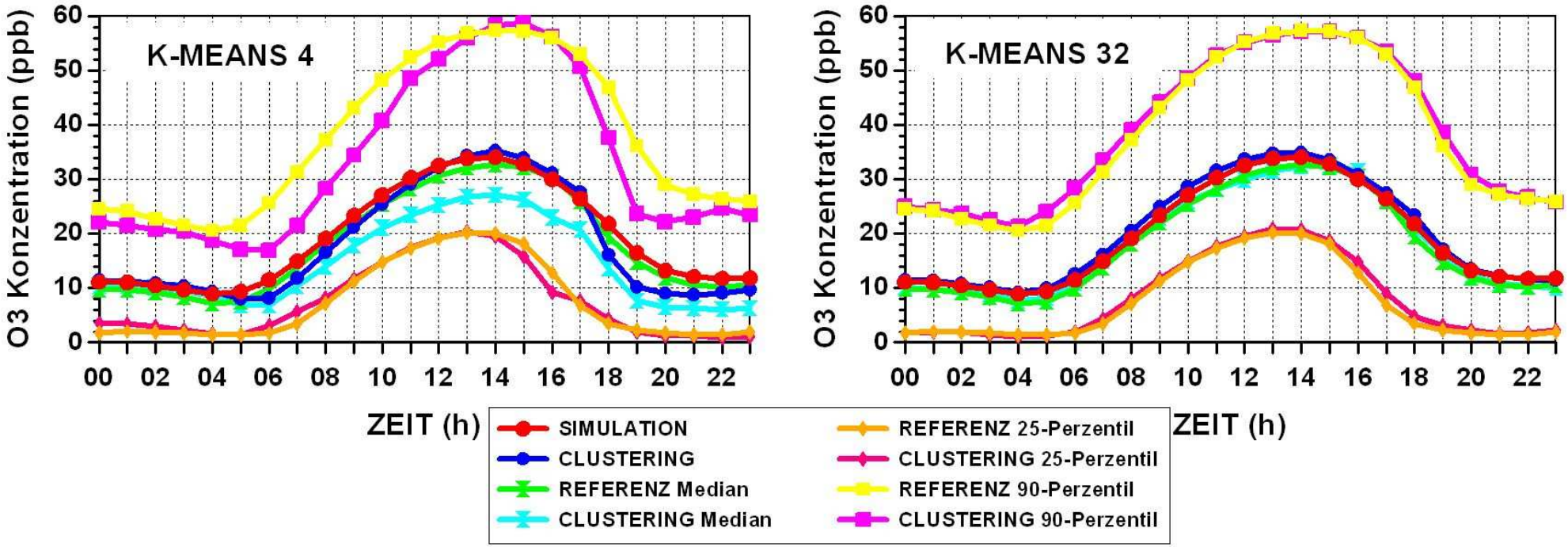
$$d_j = \sqrt{\frac{d_{jk}^2}{d_0^2}} \leq 1$$

- d_j : **standardisierte Distanz des Objektes j vom Zentrum des Cluster** g_k
- $d_{jk}^2 = \sum_{i=1}^{N} \left(x_{ji} - \overline{x}_{g_k i} \right)^2$: **für jedes Objekt j, j = 1,…., n_k, die Summe über alle quadr. euklid. Distanzen innerhalb eines Clusters** g_k
- $\overline{x}_{g_k i} = \frac{1}{n_k} \sum_{j=1, j \in g_k}^{n_k} x_{ji}$: **Klassenzentrum (Mittelwert) der Variablen i im Cluster** g_k , i = 1, N,
- $d_0^2 = \frac{SQ_{tot}}{n}$, **mittlere quadr. euklid. Distanz aller Objekte vom Zentrum der 1-Clusterlösung**
- $SQ_{tot} = \sum_{j=1}^{n} \sum_{i=1}^{N} \left(x_{ji} - \overline{x}_i \right)^2$: **Gesamtstreuungsquadratsumme;** x_{ji} : **für Objekt j, j= 1,..., n, der Wert der Variablen i, I = 1, …, N,** $\overline{x}_i$: **Mittelwert von i, i = 1, …, N, gemittelt über alle n Objekte**

K Clusterlösung erfüllt als erste Kriterium für alle Objekte. Für Lösung K' > K auch immer erfüllt.

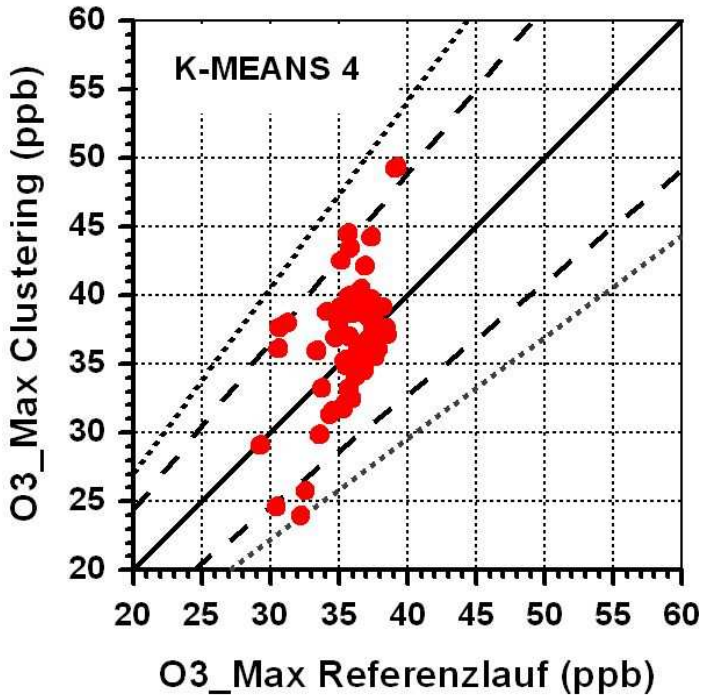
Sensitivität der Ergebnisse bei unterschiedlicher Clusterzahl:

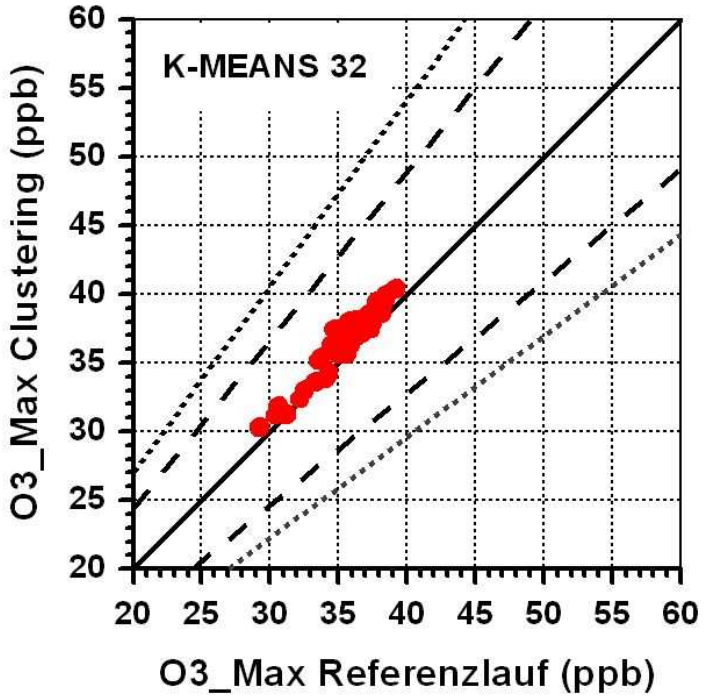
Beispiel Parameterkombination 1
Klassifizierung über das ganze Jahr
Methode: K-MEANS
Zahl der Cluster: 20 (Homogenitätskriterium)
Zahl der Cluster: 32 (Kriterium: Γ- Index)
Zahl der Cluster: 4 (Prozentuale Verbesserung gegenüber vorausgehender Clusterlösung)

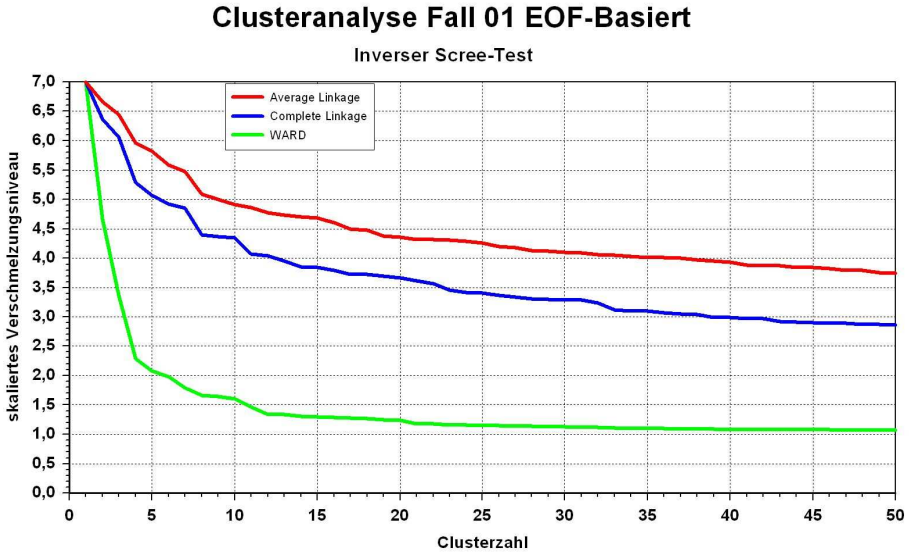


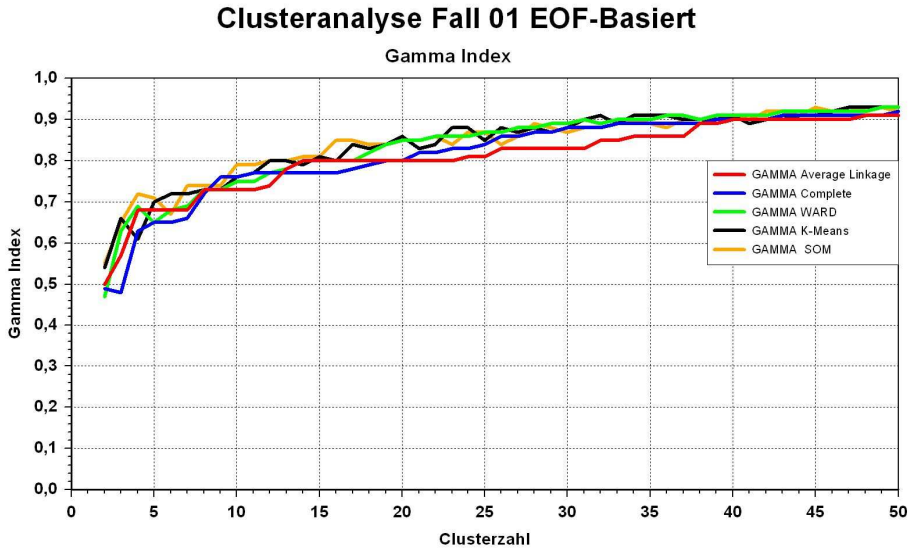
Sensitivität der Ergebnisse bei unterschiedlicher Clusterzahl:

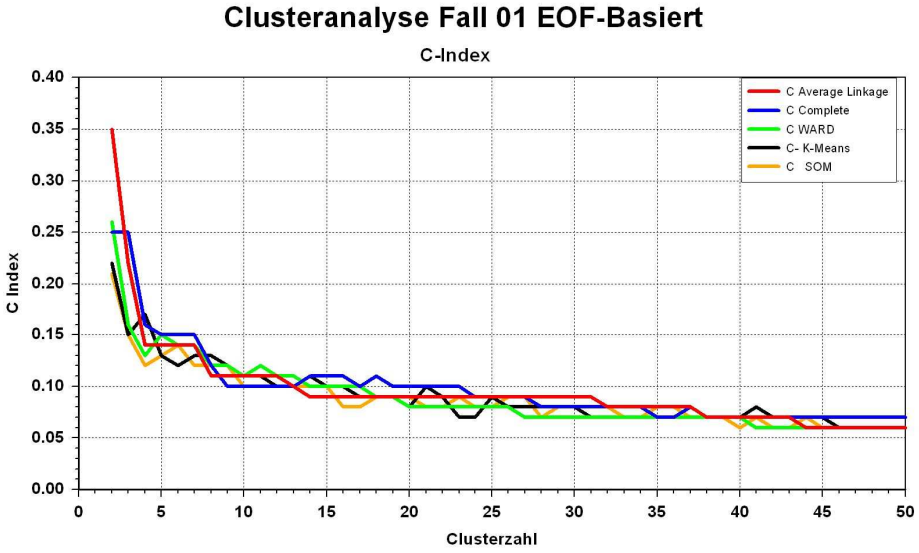
		Mittelwert (ppb)	mittl. Abweichung zur Referenz (ppb)	mittl. rel. Abweichung zur Referenz (%)
O3	Referenz	19,31	0,00	0,00
	K-MEANS 20	19,10	-0,21	-0,68
	K-MEANS 32	20,08	0,77	1,96
	K-MEANS 04	17,96	-1,35	-3,63
O3 Max	Referenz	35,64	0,00	0,00
	K-MEANS 20	36,25	0,61	0,85
	K-MEANS 32	36,53	0,89	1,22
	K-MEAN -04	36,67	1,04	1,08

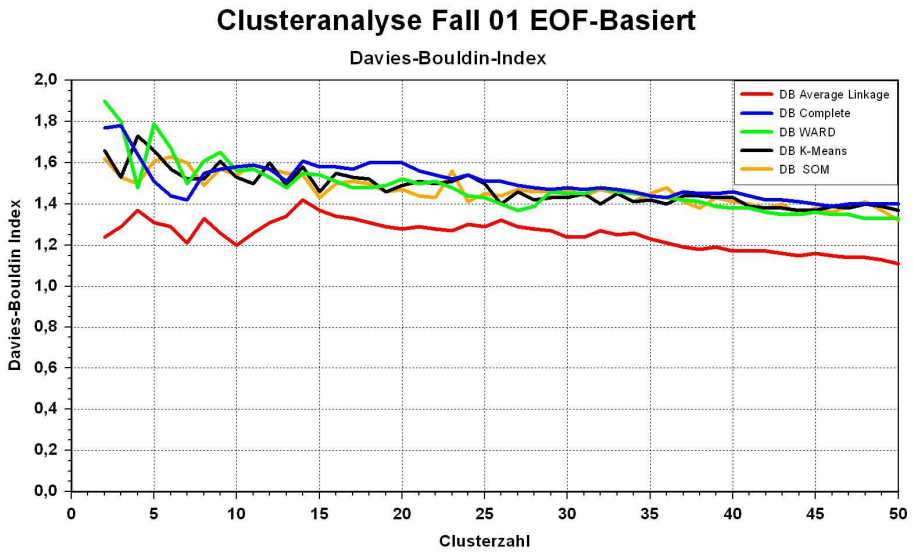


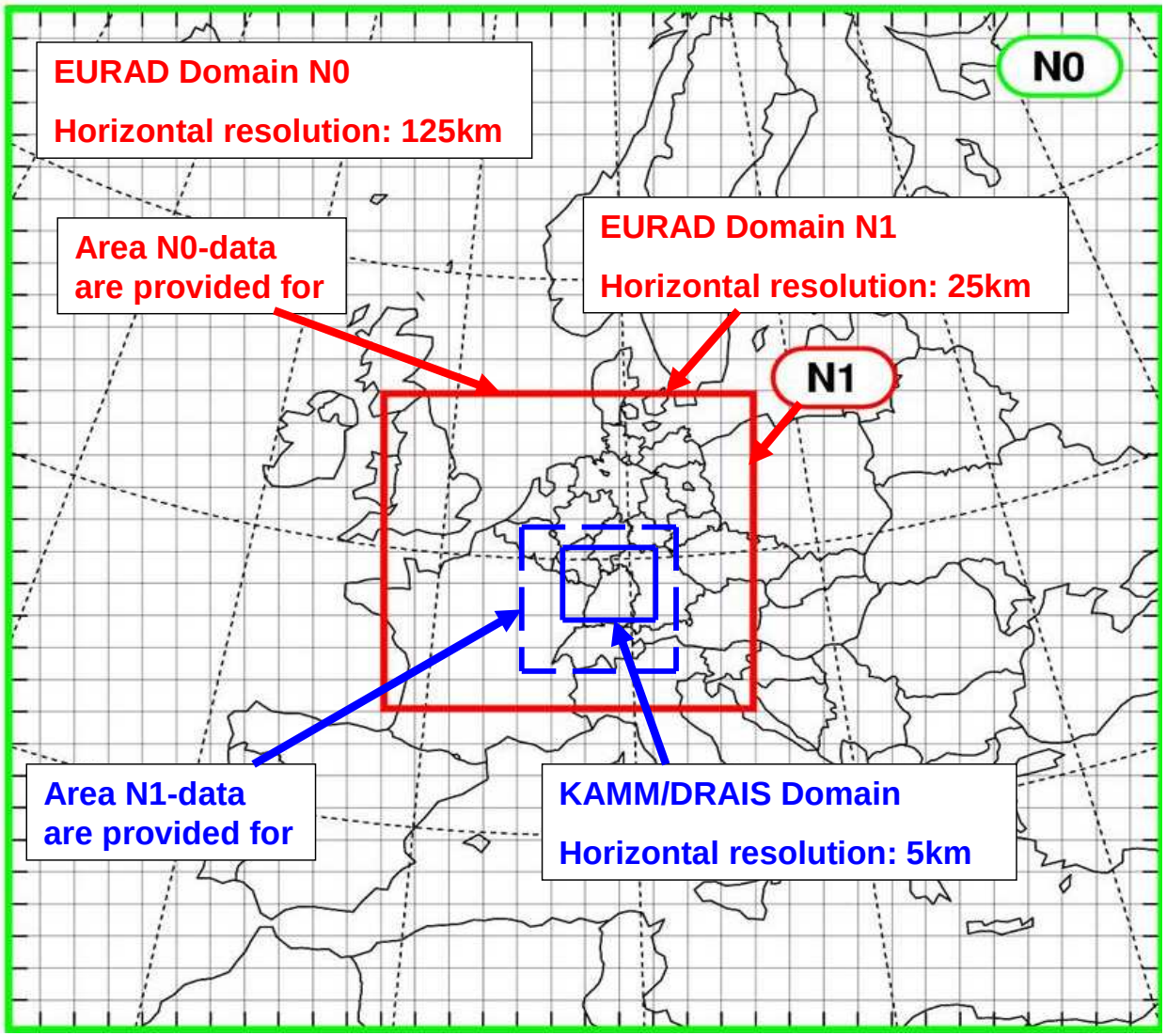



Clusteranalyse Fall 01 EOF-Basiert
Inverser Scree-Test
Average Linkage
Complete Linkage
WARD
skaliertes Verschmelzungsniveau
Clusterzahl


Clusteranalyse Fall 01 EOF-Basiert
Gamma Index
GAMMA Average Linkage
GAMMA Complete
GAMMA WARD
GAMMA K-Means
GAMMA SOM
Gamma Index
Clusterzahl


Clusteranalyse Fall 01 EOF-Basiert
C-Index
C Average Linkage
C Complete
C WARD
C- K-Means
C SOM
C Index
Clusterzahl


Clusteranalyse Fall 01 EOF-Basiert
Davies-Bouldin-Index
DB Average Linkage
DB Complete
DB WARD
DB K-Means
DB SOM
Davies-Bouldin Index
Clusterzahl


EURAD Domain N0
Horizontal resolution: 125km
N0
Area N0-data
are provided for
EURAD Domain N1
Horizontal resolution: 25km
N1
Area N1-data
are provided for
KAMM/DRAIS Domain
Horizontal resolution: 5km