Применение совместного статистического анализа термохалинных и плотностных полей для фильтрации и представления массивов данных морских гидрологических наблюдений.

Для подготовки входного потока данных в численных экспериментах предполагается использовать метод фильтрации систематических и случайных ошибок в массивах данных гидрологических наблюдений [7] . Фильтрация проводится в два этапа: вначале на каждом горизонте на основе анализа эмпирических гистограмм совместных функций плотности вероятности (ФПВ) *T-S-ρ* значений на каждом горизонте с участием эксперта фильтруются систематические и случайные ошибки данных, на втором этапе фильтруются те гидрологические станции, на которых процент отбракованных данных на горизонтах превосходит установленный экспертом пороговый уровень. Разработанная в методе технология автоматизированного построения ФПВ *T-S-ρ* гистограмм реализуется с учётом требуемых в расчётах масштабов пространственно-временного осреднения и наличия в данных наблюдений неоднородностей в их распределении в пространстве и времени. Погрешности в данных, имеющих случайную природу, удаляются из рядов наблюдений автоматизированным путём с учётом задаваемого экспертом доверительного интервала вероятности. Предлагаемая методика фильтрации ошибок применима для любых видов ФПВ *T-S-ρ* значений, включая многомодальные формы в их распределении.

Рассматривается задача фильтрации ошибочных данных гидрологических наблюдений в конкретном географическом регионе. Исследуемый массив данных состоит из дискретных значений температуры – *T*, солёности – *S* и рассчитанным по этим данным плотности морской воды – *ρ*. Перечисленные данные в этом массиве могут быть неоднородно распределены в пространстве и времени.

Для простоты начального изложения предлагаемого метода делается допущение об отсутствии в распределении анализируемых наблюдений пространственно-временных неоднородностей. По этой же причине опускаются подробности, связанные с масштабами их пространственно-временного осреднения. В этих условиях статистические веса – δ рассматриваемых наблюдений полагаются одинаковыми и равными единице. Перечисленные выше важные, но не принципиальные для излагаемого метода обстоятельства будут учтены и введены в расчёты далее на этапе обобщения предлагаемого метода для его применения в реальных природных условиях.

Пусть в пределах исследуемого региона на анализируемом гидрологическом горизонте имеется N одновременно измеренных значений *T*, *S* величин, для которых определены соответствующие им значения плотности - *ρ*. По этим данным в пространстве *T-S* координат строится сеточная область с шагом по координатам *T* и *S*, определяемыми формулами:

$\Delta T = \mathrm{Max}((T_{\max} - T_{\min})/(N-1), \mathrm{d}T)$,

$\Delta S = \mathrm{Max}((S_{\max} - S_{\min})/(N-1), \mathrm{d}S)$,

где d*T* и d*S* инструментальная точность определения T и S значений. Далее проводится процедура подсчёта количества попаданий в каждую ячейку построенной сеточной области данных из анализируемого ряда *T-S* наблюдений. После завершения этого подсчёта полученные числа

попаданий наблюдений в каждую ячейку (далее частоты попаданий или просто частоты) проходят процедуру нормировки (итоговая сумма частот по всем ячейкам после нормировки должна быть равна 1). Далее нормированные значения частот умножаются на 100%. В результате проведённых преобразований модифицированные частоты определяют в процентах частоту попадания данных в каждую ячейку на *T-S* плоскости. Введение третьей нормальной к *T-S* плоскости координаты с процентной шкалой частот позволяет построить в таком трёхмерном пространстве уровенную поверхность, которая каждой ячейки *T-S* плоскости ставит в соответствие выраженную в процентах частоту попадания в неё наблюдений. В итоге такого построения получается двумерная гистограмма ФПВ на плоскости *T-S* значений. Если же построенную поверхность уровня спроецировать на сеточную область плоскости *T-S* координат и далее с помощью алгоритма двумерной интерполяции от ломаных ступенчатых проекций уровня перейти к их сглаженным значениям, то получим двумерный сглаженный аналог описанной выше гистограммы, схематичный вариант которой представлен ниже на Рис. 6.

На этом рисунке сплошными линиями на *T-S* плоскости проведены, выраженные в процентах изолинии равных частот попадания наблюдений в диапазоны *T-S* значений. Пунктирными линиями отмечены места прохождения изолинии равных плотностей воды – ρ. Цифрами 1, 2, 3, 4, 5 помечены специально выделанные кластеры, а так же изолинии плотности, совпадающие с вершинами этих кластеров. Введённые обозначения понадобятся для наглядности демонстрации получаемых ниже обобщений.

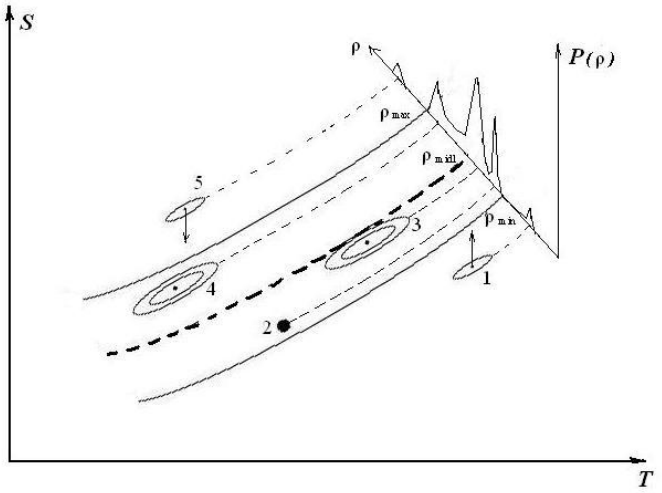

Рис. 6. Двумерная гистограмма ФПВ температуры -*T* и солёности -*S*

Ниже будут использованы следующие определения:

- область реализованных состояний – область ненулевых частот,

- область вероятных состояний – область, внутри которой экспертом удалены наиболее явные систематические ошибки наблюдений.

Из анализа Рис. 6 с учётом введённых определений следует:

- В поле силы тяжести на фиксированном горизонте диапазон вероятных состояний *T-S* значений достаточно узок. Он располагается внутри изогнутой в сторону температурной координаты полосы, верхняя и нижняя границы которой совпадает с местом прохождения максимальных- ρ_{max} и минимальных - ρ_{min} значений плотности. Вблизи центра этой полосы проходит изолиния средней на горизонте плотности – ρ_{midl}. Данные, расположенные выше и ниже указанной полосы, имеют соответственно большую или меньшую плотность. В итоге получается, что присутствие водных объёмов на исследуемом горизонте, имеющих *T-S* значения вне пределов указанной полосы, маловероятно потому, что под действием архимедовых сил они будут смещаться по вертикали соответственно или в сторону нижних или верхних горизонтов. На Рис. 6 направления этих перемещений указаны стрелками: данные, образующие кластер 5 будут опускаться, а данные кластера 1 будут пониматься на верхние горизонты.
- Кластерные квазиэллиптические образования на гистограмме соответствуют положениям водных масс, на исследуемом горизонте. Наиболее стабильны на этом горизонте те водные массы, вершины которых наименее удалены от центральной линии области вероятных состояний, которые обычно совпадает или располагается вблизи изолинии ρ_{midl}. Кластеры под номерами 1 и 5 располагаются за пределами вероятных состояний и должны быть удалены из дальнейшего рассмотрения, то есть отфильтрованы, поскольку под действием архимедовых сил все равно они должны покинуть анализируемый горизонт.
- Данные, расположенные в области точечного кластера (кластер под номером 4) требуют дополнительной проверки (возможны повторения одних и тех же наблюдений в разных массивах данных, полученных из различных источников).

Если же по описанной выше методики провести построение гистограммы ФПВ для плотности - ρ, то её вид будет подобен форме, представленной ниже на Рис.7.

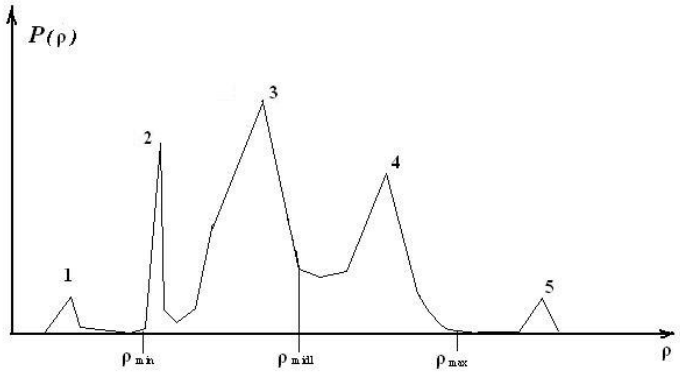

Рис.7. Гистограмма ФПВ плотности морской воды – ρ.

Из проведённого анализа большого количества гистограмм, подобных Рис.7, построенных по данным наблюдений, размещённых на различных горизонтах для различных регионов океанов и морей, следует, что:

- Гистограмма может иметь один или несколько экстремумов. В случае одного экстремума, его вершина располагается вблизи точки ρ_{midl} , если таких экстремумов несколько, то они располагаются по обе стороны от точки ρ_{midl}. Область вероятных состояний плотности на исследуемом горизонте располагается по обе стороны от точки ρ_{midl} в области, где значения ФПВ на этой гистограмме не приобретают нулевых значений.
- Маловероятными для исследуемого горизонта являются области реализованных состояний, расположенные вначале и в конце гистограммы, которые отделены от основной (центральной) области реализованных состояний (включающей в себя ρ_{midl}) нулевыми значениями ФПВ (в данном случае вершины этих областей отмечены цифрами 1 и 5). Водные массы с этими значениями плотности будут смещаться под действием архимедовых сил по вертикали вверх (1) и вниз (5) соответственно.

Для получения более полного представления о состоянии *T*-*S*-ρ данных на исследуемом горизонте на гистограмме Рис.6 в верхней правой части этого рисунка помещена гистограмма Рис.7. В представленной комбинации хорошо прослеживается однозначная связь между экстремумами этих рисунков, а так же связь изолиний ρ_{min}, ρ_{midl}, ρ_{max} (Рис.6) с соответствующими точками графике Рис.7.

Процедура фильтрации массивов наблюдений проводится в три этапа. На первом этапе из анализируемых рядов наблюдений экспертом удаляются систематические погрешности. Для этого ему предлагаются для анализа построенные в автоматическом режиме описанные выше гистограммы (Рис.6,7). Эксперт выделяет на них вызывающие сомнения структурные образования (кластеры под номерами 1, 5 и 4), определяет их границы и с помощью автоматизированных технологий данные наблюдений, образующие эти кластеры выделяются из анализируемых массивов. Далее экспертом принимаются решения по поводу целесообразности удаления этих наблюдений из дальнейшего рассмотрения.

Последнее действие эксперта на этом этапе – определение и ввод в систему для дальнейшей автоматизированной обработки граничных значений ρ_{min} и ρ_{max}, определяющих границы зоны вероятных состояний наблюдений. Численные значения этих величин экспертом определяются из визуального анализа упомянутых выше гистограмм и вводятся им в систему вручную, в качестве управляющих параметров. После установления границ зоны вероятных состояний происходит автоматическое удаление (фильтрация) наблюдений, расположенных на *T*-*S* плоскости за пределами этой зоны.

Второй этап фильтрации связан с удалением случайных погрешностей в данных, расположенных в зоне вероятных состояний в пределах задаваемого экспертом уровня доверительного интервала вероятности. Алгоритм решения этой задачи при наличии двумерной

ФПВ на *T-S* плоскости является стандартной вычислительной операцией, и включён в реализованную в данном методе расчётную технологию. Поскольку в рамках излагаемого в работе материала он не представляет самостоятельного интереса, здесь он не обсуждается.

Третий этап фильтрации наступает после реализации первых двух её этапов на всех рассматриваемых горизонтах. В нём отбираются и удаляются из рассмотрения те гидрологические станции, на которых количество горизонтов с отфильтрованными данными превосходит установленный экспертом предел.

Таким образом, в методе завершается процедура объёмной фильтрации наблюдений, а оставшиеся после трёх этапов фильтрации данные образуют массивы фильтрованных наблюдений.

Предложенный метод допускает обобщение на случай наличия в наблюдениях неоднородности в их распределении в пространстве и времени, а так же влияния на расчёты орографии рельефа дна. Последняя может изменять водную площадь сечений морской поверхности на различных горизонтах, а, следовательно, влиять на расчёты плотности распределения данных наблюдений на единицу площади водной поверхности. Учёт неоднородности в распределении наблюдений проводится с применением требуемых масштабов пространственного и временного осреднений. Используемый при этом принцип заключается в уменьшении статистических весов тех данных, плотность распределения которых в пространстве и времени превосходит среднее значение.

Реализация этой логики осуществляется следующим образом. Статистический вес наблюдения - δ на рассматриваемом фиксируемом горизонте *Z=D*, при общем количестве этих наблюдений - Q, определяются формулой (6):

$$\delta = \delta_L \delta_\tau / C\,, \qquad (6)$$

где δ_L и δ_τ соответственно весовые пространственные и временные множители, а C – нормировочный множитель, определяемый выражением:

$$C = \sum_{i=1}^{i=Q} \delta_i \qquad (7)$$

Для определения множителя δ_L исследуемая область океана на рассматриваемом горизонте разбивается на сеточную область. Шаги этой сетки в направлении осей широт и долгот равны требуемому в постановке задачи осреднению в направлении этих координат. Выбирается единица измерения площади, после чего для каждой ячейки сеточной области определяется своя плотность размещения на ней данных наблюдений – γ. Далее определяется средняя по всему горизонту плотность данных наблюдений, приходящаяся на единицу его водной площади – $\gamma_{ср}$. Величина весового множителя – δ_L, каждого наблюдения определяемого формулой (8):

$$\delta_L = \begin{cases} 1, & \gamma \le \gamma_{ср} \\ \gamma_{ср}/\gamma, & \gamma > \gamma_{ср} \end{cases}, \qquad (8)$$

где γ плотность данных наблюдений в той ячейке, где оказалось это наблюдение.

Аналогичным образом определяется временной весовой множитель. Итоговый временной интервал T, в котором размещаются анализируемые данные, разделяется на временные отрезки, равные масштабу временного осреднения – τ,. Для каждого из полученных временных отрезков определяется своя приходящаяся на единицу времени плотность данных наблюдений – η. Далее определяется средняя плотность данных наблюдений в пределах всего интервала наблюдений – $\eta_{ср}$. Затем временные весовые множители наблюдений – δ_τ определяются представленной ниже формулой:

$$\delta_\tau = \begin{cases} 1, & \eta \le \eta_{ср} \\ \eta_{ср}/\eta, & \eta > \eta_{ср} \end{cases}, \qquad (9)$$

где η плотность данных наблюдений на временном отрезке, где располагается анализируемые данные наблюдений.

Применение предложенной технологии позволит получить результаты, на которых более чётко и рельефно выделяются отдельные структурные образования. В отличии от ранее опубликованных методов похожего типа, в нем отсутствуют ограничения, связанные с формой ФПВ анализируемых массивов наблюдений, а их анализ осуществляется на основе совместного анализа массивов *T-S-ρ* значений, что ранее так же не реализовывалось. Численная реализация предложенной технологии за счёт автоматизации технических работ качественно уменьшает временные затраты, связанные с фильтрацией данных наблюдений, а удобное представление для анализа результатов статистической обработки рассматриваемых массивов наблюдений улучшает качество экспертных оценок. Отсутствие ограничений на форму ФПВ позволило автоматизировать технологию фильтрации случайных погрешностей. Технология адоптирована к реальным ситуациям, поскольку в ней учитывается пространственно-временная анизотропия в распределениях наблюдений и все расчёты проводятся в требуемых масштабах пространственно-временного осреднения. В некотором смысле, данная работа аналогична работе [7], но там рассматривалась задача селективного сглаживания гидрологических полей, а не задачи фильтрации массивов данных.

Следует так же отметить, что гистограммы совместных ФПВ на *T-S* плоскости, построенные по фильтрованным данным термохалинных наблюдений, могут быть использованы в качестве тестов, для проверки результатов численного моделирования процессов, происходящих в океане в конкретном исследуемом регионе. Для этого требуется создать по результатам моделирования на каждом горизонте соответствующие массивы расчётных значений *T-S-ρ*, построить по ним гистограммы и провести их сопоставление с гистограммами, полученными по фильтрованным натурным данным. Качество совпадения этих гистограмм позволит судить о качестве численного моделирования океанических процессов. Такое сопоставление будет физически более обоснованным, нежели сопоставление карт средних значений моделируемых полей с картами средних значений этих же характеристик, построенных по данным наблюдений,

поскольку карты этих характеристик в природе никогда не наблюдались, а являются технологичным продуктом самих исследователей.